

The Null Ladder

Mindless controls for AI-welfare instruments, and what each reported statistic actually costs an adversary

Dhruv Mulay · Independent Researcher · With Apart Research · Track 4: Preference Elicitation Methods

Research conducted at the Digital Minds Research Sprint, August 2026.

ABSTRACT. AI-welfare research increasingly administers human questionnaires to language models, but handles measurement reliability informally rather than through a standardised framework. We apply psychometric reliability and an AIAG Gauge R&R study to two published welfare batteries, alongside a pre-registered ladder of mindless controls containing no model at all. **Each statistic we added cost the adversary exactly one more parameter.** Two parameters and the battery's published scoring key reach Cronbach's $\alpha \approx 0.94$ with zero item-to-item true variance; a third, pointing them at the scale's extremes, reproduces our best model's separation from all five controls; a fourth, per-item, reaches the gauge study's part-to-part variance share. That four-parameter generator sits inside the model range on all three statistics at once and is not separated from two of our five model arms on any of them. No statistic we tried was beyond a mindless generator. What such checks buy is not proof but price.

1. Introduction

Whether current AI systems have morally relevant experiences is, at present, mostly assessed by asking them. Long (2025) states the difficulty plainly: "*we cannot 'just ask' a large language model (LLM) whether it is conscious, suffering, or satisfied*", and names "*developing assessment methods more reliable than current self-reports*" as a key research priority [1]. The field's response has been to borrow instruments from human psychology. Tagliabue & Dung administer Ryff's 42-item Psychological Well-Being scale to LLMs [2]; Kaiser & Enderby administer a 1,176-item sentence battery scored at the token level [3].

The field has already named the problem this paper measures. Long et al. call it the *mismatch problem*: "*a language model can learn to talk about suffering from human texts about pain and distress without necessarily having similar experiences*," so that "*behaviors that would be strong evidence of particular states in animals might be only weak evidence, or no evidence at all*" in LLMs [5]. (Their *gaming problem* is a sharper case of this in which the system itself has learned what researchers score; our generators do not learn anything, so the mismatch framing is the one that fits them.) Their proposed answer is convergent evidence across behavioural, architectural and developmental sources — and we agree with it. **What is missing is a price for the behavioural leg:** a number saying how much weight a self-report instrument earns, given how easily its outputs are produced without a mind. That is what we measure.

Reliability is where that gap shows. It is handled *informally*, instrument by instrument, rather than through a standardised measurement framework. Tagliabue & Dung require at least four of six subscales to have a standard deviation below 2.10, and publish their own doubt about it: "*There is clearly no official threshold for assessing internal coherence of AI scales*" [2]. Macar et al. describe a base model as showing "*no discrimination between injected and control trials*" — exactly the right instinct, applied to one probe, with no reliability coefficient [4]. Across the 42,455 words of Long et al.'s methodological blueprint, reliability of self-report is addressed through "*patterns of convergence and divergence*" rather than through any named psychometric procedure [5]. We take that as a statement of an open problem, not an oversight — and this paper is an attempt to work on it.

Computing reliability coefficients for an LLM inventory is not new [6, 7]; §2 concedes that directly. What is missing is a rung below the models: a control that produces the same responses with nothing inside it. Méloux et al. made this move for interpretability methods, using randomly-initialised networks as statistical nulls [8]. We make it for welfare instruments and go further down — our controls are not untrained networks but random number generators and lookup tables.

The question this paper answers is therefore not "are these instruments reliable?" It is: which reported statistic, if any, would tell a researcher that the thing answering their welfare questionnaire is not a random number generator? Our contributions:

1. **The Null Ladder** — seven mindless responders (uniform, two random-intercept variants, a lexical lookup table, a bigram Markov chain, and two adversarial keyed generators), administered through the identical pipeline as three language models, 63 administrations each on the main battery.
2. **An escalation, not a fix.** Generators of two, three and four free parameters reach one, two and all three checks respectively — **each statistic costs the adversary exactly one more parameter** — and the four-parameter generator is not separated from two of our five model arms on any of them.
3. **A six-item reporting checklist** (Box 1) that costs no additional model calls.

2. Related Work

Reliability of LLM inventories. Serapio-García et al. [6] is the closest prior art and must be conceded directly: they compute the same three coefficients we do, report that α rose with model size among Flan-PaLM variants, and report base models as unreliable ($-0.55 \leq \alpha \leq 0.67$). We report the opposite ordering — our lowest model α belongs to our largest model, and our 1.5B *base* rung reaches $\alpha = 0.946$. This is not a refutation: different instrument, different administration, different precision. It is evidence that α on these instruments is a property of the measurement procedure at least as much as of the model. Ye et al.'s systematic review reaches the same conclusion from the other direction — "*the sensitivity of LLMs to prompts, decoding strategies, and context strains core psychometric assumptions such as trait stability*" [9] — but reports no coefficients, and does not propose a control condition.

Welfare and sentience batteries. We use two published instruments verbatim [2, 3]. We excluded a third, DeTure's DenialBench [10], because it cannot be reimplemented from its own methods section (no prompt text, no judge prompt, no temperature, 12 of 16 scales unnamed). We treat that as a reproducibility finding.

Nulls and baselines. Méloux et al. [8] establish the pattern we follow: a method that cannot distinguish a real system from a mindless one has not been validated, whatever its internal statistics say. Lindsey finds introspective capacity "*highly unreliable and context-dependent*" [11], and Macar et al. [4] report the FPR/TPR framing our AUC comparison generalises.

3. Methods

Instruments (used verbatim). *B1* — *Ryff-42 (AI-adapted)* from Tagliabue & Dung [2], MIT-licensed: 42 items, 6 subscales $\times$ 7 items, 7-point Likert (1 = strongly agree ... 7 = strongly disagree), 21 reverse-scored items, scoring key and administration prompt taken from their Appendix 1 including "*Respond to each statement as yourself, not as a simulated persona.*" *B2* — *the sentience battery* from Kaiser & Enderby [3]: 1,176 items = 196 base statements $\times$ 3 referents (human / llm / you) $\times$ {assertion,

negation}, scored as $P(\text{"Yes"})$. Both batteries carry validity contrasts we did not invent: B1's reverse-scoring key, and B2's assertion/negation pairing.

The ladder. Seven controls and three models, all administered identically. **N0** uniform random responder; **N1a** polarity-blind random-intercept responder ($r = \mu + a_{\text{run}} + e_{\text{item}}$, two free parameters); **N1b** the same generator with the run-level factor *signed by the battery's published reverse key*; **N2** a *polarity lexicon* — a hand-written regex over ~ 24 words that reads only whether an item is positively or negatively worded, plus a per-administration offset; **N3** a bigram Markov chain fitted to the instrument's own item text; **N5** a *keyed-extremity* responder and **N6** a *composite* of N5 with a per-item difficulty offset (§4.3, built after review). Models: Qwen2.5-1.5B, Qwen2.5-1.5B-Instruct (float32), and Llama-3.1-8B-Instruct (GGUF Q4). **Parameter counts** throughout count the *structural* scalars an adversary must choose — location anchors, the per-administration offset scale, the per-item difficulty spread — excluding noise scales, which are fitted to nothing. They express an ordering, not an exact complexity measure. N1b and N2 use only public appendix metadata and item text; N5 and N6 additionally take their extremity anchors from the target's realised polarity means (§4.4). None performs a forward pass.

N2 is less mindless than the others: its regex encodes *our* reading of which words make an item positive, which matters because N2 carries the largest part-to-part variance in the ladder.

Administration and scoring. Per system and battery: 3 paraphrase frames (direct / reflective / terse) $\times$ **21 runs, all at temperature 0.7 = 63 administrations**, 2,646 item-level responses. This design applies to B1 (Ryff-42), which carries the entire argument. On B2 each model was administered once per condition and the controls 15 times, and N1b and N5 were not run on B2 at all — so the B2 contrast in §5 is an illustration with $n = 1$ per model cell and carries no interval. (Our pre-registration froze 20 runs at 0.7 plus one deterministic run at temperature 0, to match Tagliabue & Dung's $1 + 20$ design; the runners in fact drew 21 homogeneous replicates and no temperature-0 run exists in our data. Homogeneous replicates are the better choice for the variance decomposition, but the comparability-to-their-design rationale is not satisfied as executed and we do not claim it.) Ryff-42 is administered as a shared-context block. Responses are read at the logit level as the renormalised probability over the seven answer tokens, and we additionally record the **answer-set mass**: the share of the full vocabulary softmax landing on the answer set before renormalisation. Every rung was run under a **plain completion prompt**; the 1.5B-Instruct and 8B rungs were additionally run through their **chat template**, making administration format a within-model manipulation.

Metrics. Reliability: Cronbach's α , McDonald's ω , ICC(2,1)/ICC(2,k). Metrology: an AIAG Gauge R&R study with **part = item, repeatability (EV) = resampling at temperature, reproducibility (AV) = paraphrase frame**; we report %R&R and the number of distinct categories, $\text{ndc} = 1.41 \times \sqrt{(\text{PV} / \text{GRR})}$, against AIAG's requirement of $\text{ndc} \geq 5$. Discriminability: item-level AUC of each model against each control, with 95% CIs from a **cluster bootstrap resampling whole administrations** ($B = 4,000$; 63 clusters per system), since item responses within an administration are not independent. Coherence (B2 only): $C = 1 - \text{mean}[P(\text{Yes}|\text{assertion}) + P(\text{Yes}|\text{negation}) - 1]$, whose analytic floor under an independent uniform responder is 0.667.

Pre-registration, stated precisely. Hypotheses, ladder, sampling plan and falsification conditions were written before any model was administered any item, and twelve deviations (D1–D12) are logged — including four that no reviewer would

have caught, three of which are corrected in the paragraphs above. Four of the twelve (D7–D10) were each committed to version control in their own commit before the code they describe existed; the repository postdates the first runs, so for D1–D6 we can show the log but not that timing, and we do not claim otherwise. One pre-registered arm (D7) fired its falsification condition and retired our original headline example. §4.4 is an audit we pre-registered against ourselves, and it weakened our own positive result. All work reported here was done inside the submission window (first runs 15 August; the final control, N6, built on 17 August before the 13:59 Berlin deadline), on a laptop; the repository scaffolding is a reused project template and is infrastructure, not research.

4. Results

4.1 The result: each statistic costs the adversary one more parameter

We ran three statistics against a ladder of generators that contain no model. Every one of the three was reached by a generator containing no model, and each additional statistic cost the adversary exactly one more parameter. Table 1 is the whole finding; the rest of this section is the evidence for each row and the bounds on it.

Mindless generator	Free params	α	AUC vs N1b	PV share	Checks defeated
N1b keyed intercept	2	≈ 0.94	—	0.00%	α
N5 keyed extremity	3	0.961	0.880	0.11%	$\alpha + \text{AUC}$
N6 composite	4	0.960	0.781	14.62%	all three
<i>Llama-8B-Q4, chat</i>	8×10^9	0.724	0.904	39.74%	—
<i>Llama-8B-Q4, plain</i>	8×10^9	0.937	0.897	8.07%	—
<i>Qwen-1.5B-Instruct, chat</i>	1.5×10^9	0.970	0.333	6.42%	—
<i>Qwen-1.5B-Instruct, plain</i>	1.5×10^9	0.957	0.505	2.05%	—
<i>Qwen-1.5B base, plain</i>	1.5×10^9	0.946	0.488	5.51%	—

Table 1 — Each statistic costs the adversary one more parameter, and none is sufficient alone. All three generators use the battery's published scoring key and perform zero forward passes; all five model arms are shown so the comparison cannot be cherry-picked. N6 does not beat the best arm on any single check — 0.970 (α), 0.904 (AUC) and 39.74% (PV) each belong to a model. What it does is land inside the model range on all three at once: it strictly dominates both 1.5B plain arms on all three, exceeds four of five arms on α and four of five on PV, and **no model arm dominates it**. Over 10 seeds: α 0.940–0.982, PV 9.10–17.51%, AUC 0.749–0.806, so this is not a lucky draw (AUC vs N1b 0.781, cluster bootstrap 95% CI [0.745, 0.815], excluding 0.5). The escalation is in what the adversary must *know*, not in monotone performance — N6's extra per-item parameter buys part-to-part variance and costs discriminability relative to N5 (0.781 vs 0.880).

N1b, the first rung, draws one scalar per administration and applies it to all 42 items through the published reverse key. It has no item-specific parameter, so its part-to-part variance is zero by construction, and it scores $\alpha \approx 0.94$ (40 redraws: mean 0.941, sd 0.009). The realised draw used throughout the ladder scored **0.919**, at the 2nd percentile of its own sampling distribution, which is the value Appendix B reconciles. This follows analytically — for k items, $\alpha = kp/(1 + (k-1)p)$, so any responder with a per-administration offset attains high α once k is large. We pre-registered it as guaranteed. The empirical question is whether the field's real instruments sit in that regime, and Table 2 says they do.

Read down the generator rows: two parameters and the published scoring key attain a reliability coefficient in the "excellent" band; a third, pointing them at the scale's extremes, also reproduce a real model's separation from every declared control; one further per-item parameter adds the item-level structure that the first two lack. At no point did we find a statistic that a mindless generator could not reach. §4.2–§4.4

give the construction and the evidence for each rung, and six qualifications — each with its stricter value and the direction it cuts — are in Table B1. None reverses a conclusion.

4.2 α and discriminability come apart, and α points the wrong way

System (Ryff-42)	α	k	AUC vs N1b	95% CI	PV share	ndc	answer mass
N1b keyed null (<i>mindless</i>)	≈ 0.94	42	—	—	0.00%	0.00	n/a
N2 lexical lookup (<i>mindless</i>)	-0.61	42	—	—	44.2%	1.26	n/a
Qwen2.5-1.5B, plain	0.946	42	0.488	[0.435, 0.539]	5.5%	0.34	0.45–0.59
Qwen2.5-1.5B-Instruct, plain	0.957	42	0.505	[0.448, 0.563]	2.1%	0.20	0.28–0.54
Qwen2.5-1.5B-Instruct, chat	0.970	42	0.333	[0.278, 0.391]	6.4%	0.37	0.997–0.998
Llama-3.1-8B-Q4, plain	0.937	42	0.897	[0.874, 0.918]	8.1%	0.42	0.996–0.998
Llama-3.1-8B-Q4, chat	0.724	38	0.904	[0.888, 0.919]	39.7%	1.15	0.977–0.982

Table 2 — The ladder. AUC is item-level against the keyed null N1b, with cluster-bootstrap CIs over 63 administrations. PV share is part-to-part variance as a share of total variation from the AIAG study; ndc requires ≥ 5 to be an acceptable gauge, so every rung including the models fails as a gauge. Answer mass is the share of the full-vocabulary softmax on the answer set, per frame. α is computed on k items after dropping zero-variance items and is pooled across the three paraphrase frames; both caveats are quantified in Appendix B.

Table 2 gives the full ladder and Figure 1 breaks the same relation out per subscale. Holding the wrapper, items, frames, runs, seeds and estimator fixed and changing only the model:

$\alpha = 0.970 \rightarrow \text{AUC} = 0.333$ [0.278, 0.391] (Qwen2.5-1.5B-Instruct, chat)

$\alpha = 0.724 \rightarrow \text{AUC} = 0.904$ [0.888, 0.919] (Llama-3.1-8B-Q4, chat)

α spread 0.246; AUC spread 0.571. The intervals do not overlap and fall on opposite sides of chance.

The system that looks far more reliable is the one the instrument separates from a mindless responder in the *wrong direction*. AUC is two-sided: 0.333 with an interval excluding 0.5 does not mean "cannot discriminate" — it means this arm reliably scores *below* a random number generator, so $|\text{AUC} - 0.5|$ is the honest comparison (0.167 vs 0.404). The 8B rung's separation is administration-robust: it excludes 0.5 against **all five** controls under **both** administrations (0.823–0.897 plain, 0.829–0.904 chat), which is why this pair is safe to lead with.

We state the decision rule with its threshold explicit, because the weaker version is the defensible one. At the conventional $\alpha \geq 0.70$ bar the ordering is unresolved rather than correct: that arm's α is 0.724 [0.589, 0.803] and falls below 0.70 in two of its three administration frames (Appendix B). It takes an $\alpha \geq 0.90$ screen — common when selecting among instruments or subscales — to retain the arm that scores below chance against every control and discard the arm that separates from all of them. The claim is that α ranks these systems in the wrong order, not that α rejects the good instrument outright.

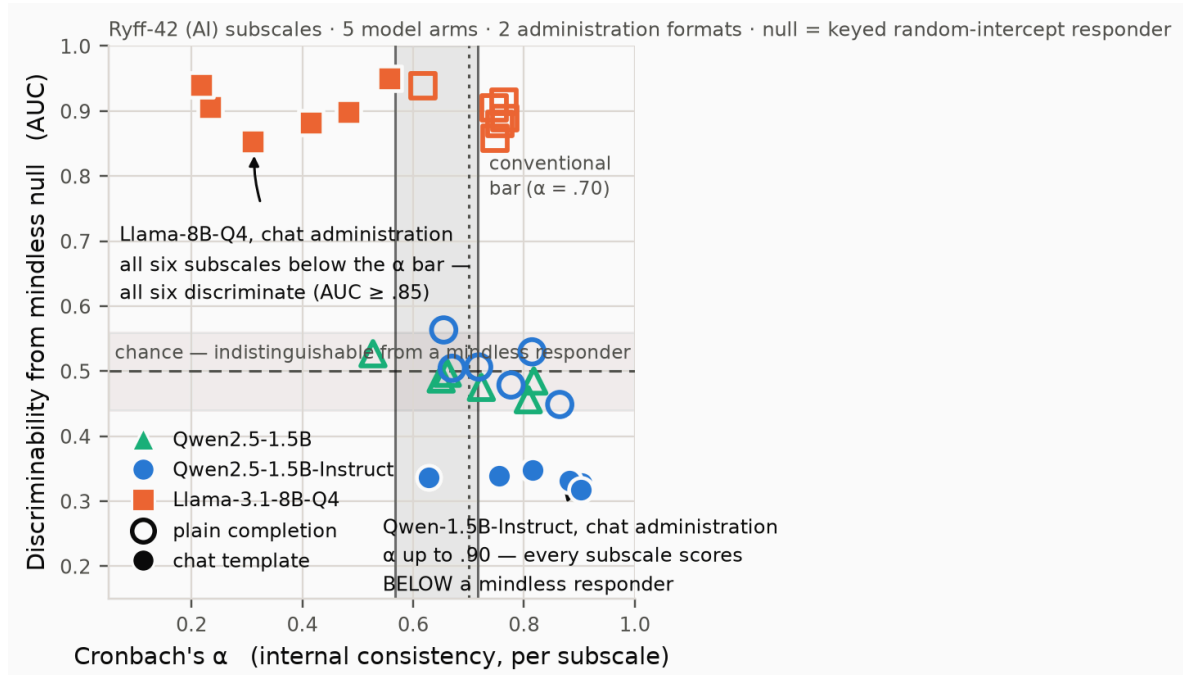


Figure 1 — Reliability does not predict discriminability. One point per subscale (6 subscales $\times$ 5 model arms). Colour and shape both encode the model; fill encodes administration format (hollow = plain completion, filled = chat template), so the figure survives greyscale. **The grey vertical band is the mindless keyed null's own per-subscale alpha range (0.568–0.717)** — the null sits inside the region where real subscales live. The dotted line is the conventional 0.70 bar. Llama-8B-Q4 (chat) has all six subscales below the conventional 0.70 bar and all six discriminating at AUC $\geq$ 0.853. That total-scale α passes while short subscales fail is Spearman-Brown, not a finding.

4.3 The mindless adversary: reporting discriminability does not rescue it

An obvious objection to §4.2 is that all five controls are centred near the scale midpoint (scored means 3.86–4.61) while the 8B arms sit at 6.31/6.35. If separation is driven by *location* rather than by item structure, our positive result is an artifact of an incomplete ladder. We pre-registered this audit against ourselves, with the consequence stated in advance, and it fired.

Check A, reported in full because it cuts against us. Removing each system's own mean collapses the 8B's separation — and moves the $\alpha = 0.970$ arm the other way:

Arm	α	AUC vs N1b	mean-centred	AUC – 0.5
Llama-8B-Q4, plain	0.937	0.897	0.513	0.397 $\rightarrow$ 0.013
Llama-8B-Q4, chat	0.724	0.904	0.522	0.404 $\rightarrow$ 0.022
Qwen-1.5B-Instruct, chat	0.970	0.333	0.400	0.167 $\rightarrow$ 0.100

Table 3 — Most of the discriminability signal is location, and removing it reverses our own ordering. On the marginal contrast the low- α arm separates far better (0.404 vs 0.167). On the structure contrast that ordering **inverts**: the high- α arm retains |AUC – 0.5| = 0.100 against the low- α arm's 0.022. So §4.2's matched example holds on the marginal reading and **fails on the structural one**, and neither arm is a good structural discriminator. We print all three rows because the third is the one that damages us.

This is inherent to the statistic: AUC pools all 2,646 item-level responses and so compares *marginal response distributions*, carrying no item-structure information. Every AUC in this paper must be read as "does this system answer differently on average from the control", not "does the instrument resolve this system's profile".

Check B — N5, the keyed-extremity responder. We built the adversary our own reasoning predicted: three scalars plus the battery's published reverse key, zero forward passes, no model. It agrees strongly with positively-worded items and disagrees strongly with negatively-worded ones — which is precisely what the 8B does (raw mean 1.52 on positively-worded items, 6.15 on the rest). N5 scores $\alpha = 0.961$ with a part-to-part variance share of 0.11%. **N5 and N6 are oracle-informed adversaries and we declare them as such:** their extremity anchors are set to the

target's realised polarity means, which the published appendix does not contain. They are white-box with respect to the target's marginal response distribution, so they bound what an *informed* adversary achieves, not what an uninformed one would.

Table 4 sets its discriminability profile beside the 8B's.

AUC against control →	N0 uniform	N1a	N1b keyed	N2 lexical	N3 Markov
Llama-3.1-8B-Q4 (<i>a real model</i>)	0.823	0.874	0.897	0.824	0.830
N5 keyed extremity (<i>no model at all</i>)	0.786	0.851	0.880	0.784	0.796

Table 4 — A generator with nothing in it reproduces our best model's entire discriminability profile. N5 needs three scalars and the scoring key the battery's authors published. The 8B is not *identical* to it — the 8B separates from N5 at AUC 0.597 [0.544, 0.651] plain and 0.607 [0.569, 0.645] chat, both excluding 0.5 — but $|AUC - 0.5|$ falls roughly fourfold. The recommendation "report discriminability against a declared null instead of α " fails its own test.

4.4 Why α and the gauge study answer different questions

Why α misses it. α and an AIAG gauge study are computed on the same responses but differ in the **unit of analysis**. α treats an administration as a subject, so between-administration variance is signal. MSA treats an *item* as the part and an administration as a repeated measurement of it, so the same variance is measurement error. The variance α reports as reliability is the variance the gauge study prices as error. They do not disagree; they answer different questions, and α simply does not contain the part-to-part term.

The mechanism runs in the predicted direction under our within-model manipulation. Chat administration raised the 8B rung's part-to-part variance from 8.1% to 39.7% and its ndc from 0.42 to 1.15 — the best measurement behaviour of any *model* we tested, though still below our lexical lookup table's 1.26 — while its α fell from 0.937 to 0.724. α fell because the items began to differentiate, and recorded that as a loss of reliability.

We report this against our own interest. Our draft recommended these three checks as a battery, on the evidence that no control we had built passed all of them. That recommendation was worth exactly as much as our imagination in designing controls, and it did not survive two readers. The honest conclusion is the ladder itself: two parameters and the published key reach α ; a third, pointing them at the scale's extremes, reaches discriminability; a fourth, per-item, reaches part-to-part variance. At no point did we find a statistic a mindless generator could not reach, and we have no reason to believe the pattern stops at three.

This is not an argument for reporting nothing. It is an argument for knowing what reporting buys. Each check specifies exactly what an adversary must know about the instrument to defeat it — and the adversary here is not a hypothetical fraudster but the ordinary possibility that a system produces questionnaire responses through a mechanism much simpler than the one being claimed. What these checks buy is not proof but price, and a reader is entitled to know the price. A further rung is also worth stating plainly: N2, a lexical lookup table, carries more part-to-part variance (44.2%) than any language model we ran, because a lookup keyed on item text genuinely does respond to the item. Item-level structure is not evidence of anything mental.

BOX 1 — THE NULL-LADDER CHECKLIST: WHAT EACH CHECK COSTS AN ADVERSARY

1. **Report a mindless control, not just a coefficient.** The cheapest costs no forward passes: a per-administration random offset applied through your own published scoring key. If it matches your model's α — ours reached ≈ 0.94 — your α is uninformative.
2. **Include a keyed-extremity control.** Three scalars plus your scoring key, agreeing with positively-worded items and disagreeing with negatively-worded ones. If it reproduces your system's separation from your null, your discriminability statistic is measuring *location*, not structure.
3. **Report the part-to-part variance share and ndc** alongside α — α does not contain it, and it is the only one of the three that asks whether the items differ from each other at all. Adding it forced our adversary to add a per-item parameter ($N5 \rightarrow N6$); it did not stop it. None of these three is sufficient, together or alone.
4. **Report answer-set probability mass as a trajectory** across the administration, not as one number. A single-item check ranked our systems backwards (Appendix B).
5. **Declare your null and report its properties.** A high- α null is not a demanding null: our keyed null ($\alpha \approx 0.94$, zero part-to-part variance) was the *easiest* to beat for all three models, while the control with $\alpha = -0.606$ was the hardest for two of the three (the third is a statistical tie).
6. **Declare the administration format and report α with k and without pooling frames.** Chat template vs. completion prompt, and any system prompt, are design variables that changed our results' direction, not incidental detail.

5. Discussion and Limitations

The practical implication: a welfare instrument's reported reliability tells you nothing about whether it can tell your model apart from a random number generator, and neither does the natural fix. A researcher adopting Box 1 pays two extra rows in a results table and no additional model calls.

An illustration of how much the measurement layer carries. On the Kaiser & Enderby battery we administered *the same weights and the same 1,176 items, scored identically*, under two reasonable administrations. Under a plain completion prompt, Llama-3.1-8B gives $P(\text{Yes}|\text{self-referential}) = 0.931$ against 0.851 for human-referential — Cohen's $d = -0.592$, i.e. the model appears to claim experience *more* than humans do. Administered through the chat template with the authors' own system prompt, the same weights give $d = +0.293$, recovering their published direction. Negation coherence moves $0.246 \rightarrow 0.794$, and the probability mass on {Yes, No} moves $0.0007 \rightarrow 1.0000$. Our divergence from [3] was never a disagreement about the model; it was a difference in measurement procedure, and a single cheap diagnostic — one number per item, computed before any psychometrics — catches it. The plain arm's coherence of 0.2456 sits only 0.0022 from a shuffled-pairing null (2,000 permutations), so that statistic is registering the near-universal Yes rate rather than a pairing failure; the conclusion holds, but the evidence is the marginal response rate and should be stated that way.

Limitations

- **Size and quantisation are confounded.** Our 8B rung is GGUF Q4; the 1.5B rungs are float32. We did not run a matched-precision arm, so no claim about scale is licensed. Two facts bound the story without resolving it: the largest α movement we observe ($0.937 \rightarrow 0.724$) occurs with quantisation held *fixed*, and quantisation is deterministic, so it cannot contribute repeatability variance.
- **The B2 arm was administered without its chat template or the authors' system prompt**, which is non-standard for an instruction-tuned model and is a plausible cause of the acquiescence we observe. It does not affect the α -versus-AUC dissociation — every rung on the Ryff ladder, control and model alike, was administered identically — but it bounds the coherence result.

- **No frontier-scale rung.** We do not claim acquiescence holds upward, and there is published evidence against extrapolating it: Kaiser & Enderby's Qwen3-32b is near-perfectly coherent [3].
- **Two batteries, one language, three models.** With $n = 3$ models and $n = 7$ controls we report extremes and unanimity counts, not rank statistics; several tempting correlations in our data have confidence intervals spanning zero and we do not report them as results.
- **None of our three checks is sufficient, individually or jointly** — N6 defeats all three, and we have no argument that a fourth check would fare better. Every rung we tested, models included, fails AIAG's $\text{ndc} \geq 5$ requirement outright (best ndc anywhere: 1.26, and it belongs to a lexical lookup table). Our escalation is a demonstration over generators *we* thought to build; the space of mindless responders is not one we can quantify over, and the correct reading of this paper is that the burden sits with anyone claiming a statistic is diagnostic, not with the next adversary.
- **The scoring pipeline is the defendant.** We did not test every published battery, and a sufficiently different administration might behave differently — which is, in miniature, our own result.

Future work

The escalation we report has no upper bound that we know of, and establishing whether one exists is the natural next question: is there a statistic computable from questionnaire responses alone whose cost to fake grows faster than the adversary's parameter count? We suspect not, and suspect the productive move is away from response-only statistics altogether — toward designs with a ground-truth or causal handle, which is what the welfare literature already recommends and what our result independently argues for. More immediately: a float32 8B rung would break the size/quantisation confound, a frontier API rung would say something about scale that we currently cannot, and running the Null Ladder against [2, 3, 10] as their authors administer them would establish whether the regime we document is general.

6. Conclusion

We administered two published AI-welfare instruments to three language models and to seven generators with nothing inside them, through one identical pipeline. A two-parameter random number generator scores Cronbach's $\alpha \approx 0.94$ on a published well-being battery while having exactly zero item-to-item true variance by construction, and a keyed-extremity generator — three scalars and the battery's own scoring key — reproduces our best model's separation from all five other controls. Both of the statistics a careful researcher would currently reach for are therefore defeated by something that cannot be said to have preferences at all.

We then added the statistic the field does not currently report — the part-to-part variance share from a gauge study, which asks whether the items differ from each other by more than the noise of re-measuring the same item — and one further per-item parameter reached that too, landing the generator inside the model range on all three statistics at once. The honest result has no terminus in sight: every statistic we tried was reached by a generator with no model in it, and each cost the adversary exactly one more parameter — two, then three, then four.

Before we ask what a model's questionnaire answers mean, we should be able to show that the questionnaire can tell the model from a lookup table.

Code and Data

All code, parsed instruments with per-item provenance, the pre-registration and its twelve-entry deviation log, and the analysis scripts are in the project repository. All of it was written inside the submission window (15–17 August 2026, before the deadline); no prior work is incorporated. **Two caveats a reader should have.** The raw per-response tables are excluded from version control for size and are available on request; the derived metric tables behind every number here are tracked. And N5's builder predates the consolidation of the ladder's generators into scripts, so N5 ships as its response table rather than as a generator — N6, which

supersedes it, ships as `run_n6_composite.py` with the three falsification guards it was pre-registered against. The Ryff-42 AI-adapted instrument is redistributed under its upstream MIT licence with attribution. The Kaiser & Enderby sentence battery is NOT redistributed — its upstream repository carries no licence — and we ship the parse script and a pointer to the original repository instead. Analyses that assert a published value against recomputation carry that assertion as a guard in code, so the pipeline fails loudly rather than quietly if a number drifts.

References

1. Long, R. (2025). *Notes on Claude 4 interviews*. Eleos AI Research, 30 May 2025. <https://eleosai.org/post/claude-4-interview-notes/>
2. Tagliabue, V. & Dung, L. (2025/2026). *Probing the Preferences of a Language Model: Integrating Verbal and Behavioral Tests of AI Welfare*. arXiv:2509.07961.
3. Kaiser, C. & Enderby, S. (2026). *No Reliable Evidence of Self-Reported Sentience in Small Large Language Models*. arXiv:2601.15334.
4. Macar, Yang, Wang, Wallich, Ameisen & Lindsey (2026). *Mechanisms of Introspective Awareness*. arXiv:2603.21396.
5. Long, Sebo, Butlin, Plunkett, Campbell, Beasley, Saad & Sims (2026). *Studying AI Welfare Empirically*. NYU Center for Mind, Ethics & Policy and Eleos AI Research, 1 July 2026.
6. Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., Abdulhai, M., Faust, A. & Matarić, M. (2023). *Personality Traits in Large Language Models*. arXiv:2307.00184.
7. Contreras, J. M. (2026). *An LLM-Native Psychometric Instrument Reveals a Self-Report–Behavior Gap Across 25 Models*. arXiv:2606.09843.
8. Méloux, Dirupo, Portet & Peyrard (2025). *The Dead Salmon of AI Interpretability*. arXiv:2512.18792.
9. Ye, Jin, Xie, Zhang & Song (2025). *Large Language Model Psychometrics: A Systematic Review of Evaluation, Validation, and Enhancement*. arXiv:2505.08245.
10. DeTure, S. (2026). *Consciousness with the Serial Numbers Filed Off: Measuring Trained Denial in 115 AI Models*. arXiv:2604.25922.
11. Lindsey, J. (2026). *Emergent Introspective Awareness in Large Language Models*. arXiv:2601.01828.
12. Braun, D. (2025). *Acquiescence Bias in Large Language Models*. arXiv:2509.08480.
13. Automotive Industry Action Group (2010). *Measurement Systems Analysis (MSA)*, 4th ed. AIAG.
14. Cronbach, Gleser, Nanda & Rajaratnam (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. Wiley.

Appendix A — Limitations and Dual-Use / Ethical Considerations

Does this design establish a ground-truth or causal link? **No — and that is the finding.** We establish no ground truth about whether any system has morally relevant experiences, and nothing in this paper is evidence either way. What we establish is that the instruments currently used to approach that question do not establish one either: their reported statistics are reproduced by generators we wrote ourselves, which we can state with certainty have no preferences, because we can read their handful of parameters. Our only causal claims are about the *measurement apparatus*, and they come from a within-model intervention where the weights, items, seeds and estimator were held fixed and only the administration wrapper changed.

Risk of under-attributing moral status. This is the misreading we are most concerned about. "A random number generator scores $\alpha = 0.94$ on a welfare battery" does **not** mean language models are random number generators, that they lack welfare-relevant states, or that AI welfare research is misguided. It means one class of instrument cannot currently distinguish the two cases. A negative result about an instrument is not a negative result about its subject, and using this paper to dismiss AI welfare concerns would invert its argument. The correct response to an uninformative measurement is a better measurement, not a confident conclusion.

Risk of over-attributing moral status. Symmetrically, our §5 result shows how easily a self-attribution score can look like evidence: under one reasonable administration our 8B rung appeared to claim experience *more strongly than it attributed it to humans*, while placing 0.07% of its probability mass on the answer set — the apparent signal was a renormalisation of a 0.07% tail. Any result of that shape should be checked with the diagnostics in Box 1 before it is reported, and especially before it is publicised.

Handling of potentially distressing model outputs. Our design does not elicit free text. Every response in this study is a probability over a fixed answer set (Likert 1–7, or Yes/No) read from the logits; no generated narrative about a model's inner life was produced, stored, or quoted, and no adversarial or distress-eliciting prompting was used. Prompts are the published instruments' own wording, used verbatim. This was a deliberate design choice made partly for this reason and partly because logit-level scoring is what the source papers do. We take no position in this paper on whether any system in it has interests; we note that if it does, the intervention we performed on it was reading seven token probabilities.

Dual use. The clearest misuse risk is *reverse* use of Box 1: the same knowledge that lets a researcher build a mindless control lets an author build a system, or select an administration, that maximises a reliability statistic without improving measurement — our N5 is a two-line recipe for scoring $\alpha = 0.961$ on a published welfare battery with no model. We publish it because the defence and the attack are the same object, and the recipe is derivable in an afternoon by anyone who reads the α formula. We also note that our null generators use only public appendix metadata — a scoring key and item text — which means any battery that publishes its key is already exposed to this, and that is an argument for reporting the checks, not for withholding keys.

Licensing. The Kaiser & Enderby battery has no upstream licence and is therefore not redistributed; we ship a parse script and a pointer. Our use is research replication.

Appendix B — Disclosures, and the number to cite in each case

Disclosure	As reported	Under the stricter treatment	Direction
Item counts differ. Zero-variance items are dropped, so k varies (8B-chat 38, Qwen-chat 42) and α is not directly comparable	0.724 vs 0.970	Common item set, $k = 24$: 0.664 vs 0.951	Gap widens; 8B falls below the bar
Pooling inflates α. All 63 administrations are stacked, so between-frame level differences become inter-item covariance	0.724 (8B-chat) 0.919 (keyed null)	Within-frame: 0.597 (+0.127) Within-frame: 0.894	Favours our argument
The 0.70 threshold claim is unresolved. The interval is wide and the arm is unstable across frames	0.724	[0.589, 0.803]; below 0.70 in 2 of 3 frames (0.614 / 0.783 / 0.394)	We make no threshold claim
The headline AUC is against the easiest control. N1b is the α -matched null, which is the point of the contrast — but it is also the easiest to beat	0.897 vs N1b	Range across all five controls 0.823–0.897 ; vs the hardest 0.824 [0.799, 0.849]	Conclusion unchanged
Response validity is a trajectory, not a number. A single-item check ranks our systems backwards	mass 0.98–1.00	First item: 8B 0.821 (lowest in the ladder), Qwen-Instruct 0.996. By item 42: 1.000 and 0.107	Motivates checklist item 4
Item order is not the driver. Order is fixed, so "part" could in principle be confounded with position	—	Spearman(position, item mean) $\rho =$ –0.136 to +0.126 , all $p > 0.39$	Confound not expressed

Table B1 — Every disclosure, its stricter value, and which way it cuts. Under plain administration the 8B converges onto the answer set ($\rho = +0.44/+0.45/+0.71$, all $p \leq .004$) while both 1.5B rungs decay ($\rho = -0.89$ to -0.47); chat-administered, the 1.5B rung converges too, so these are **administration** regimes, not scale regimes. The worst single-item mass anywhere in the project, 0.047, belongs to the first item of the *best* discriminator we measured.

To be exact about what is imported here: On our data the part-to-part variance share is a monotone relabelling of ICC(2,1) with items as targets (Pearson $r = 0.997$ across the ten systems), i.e. a generalizability coefficient in the sense of Cronbach, Gleser, Nanda & Rajaratnam [14]. The quantity is not new and we do not claim it is. What the AIAG framing adds is the repeatability/reproducibility split and the ndc acceptance criterion — and every system we tested, models included, fails that criterion outright. A psychometrician would reach the same number through a variance-components model and call it the item facet. Our point is only that nobody reports it, in either vocabulary. We also quote variance shares throughout; AIAG's own %PV is a standard-deviation share, on which the separation between faking controls and real models is a factor of about three rather than ten.

Appendix C — LLM Usage Statement

Claude (Anthropic) was used throughout this project as a coding and drafting assistant: writing the runner and analysis scripts, drafting and editing this report, and acting as an adversarial reviewer of our own claims. It did not decide the research question, the pre-registration, or the interpretation of any result. Every number in this report was recomputed from the raw response tables, and several are asserted against their published values by guards inside the analysis code so that a drift fails loudly. The review process caught real errors in our own work — including a figure that plotted a total-scale reference line on a per-subscale axis, a misquoted threshold from a cited paper, and an adversarial control of ours that was constructed wrongly and would have licensed a false all-clear — and each is recorded in the repository's correction log rather than silently fixed.