

Intensity Is Not Identified

A locked test on released model files that separates persona masks from changed rankings

Finomo Awajjogak Orom

Independent researcher¹

Digital Minds Research Sprint, 14-16 August 2026

Apart Research x NYU Center for Mind, Ethics & Policy x Eleos AI Research x CIMC

Track 4 (primary) x Track 5

Abstract

A preference-intensity number from the default assistant is not identified. Three processes produce that single-method cell: Mask (order holds, size shrinks), Different judge (order moves), and Broken items (methods disagree). We lock a sequential classifier before scoring and apply it to released, hash-verifiable files: experienced utility and self-report, default versus neutral prompt, eight models, 500 shared experiences (HuggingFace mmazeika/wellbeing-results). Two small models are Broken (pair-agreement 0.45-0.47). Six larger models are Stable. Four of eight ship byte-identical self-report "neutral" files. Decision-utility files share zero IDs with this bank. This is not a test of consciousness. Do not quote intensity until two methods on the same IDs, and two distinct voice files, have been run.

New work done during this research sprint (required disclosure)

Prior work we build on, and do not claim: Mazeika et al. (2025) Utility Engineering pairwise utilities and the public mmazeika/wellbeing-results dump; Anthropic (2025) on model welfare; Shanahan, McDonell, and Reynolds (2023) on assistant role-play; nostalgebraist (2025) on the constructed assistant. **What is new this weekend:** (i) a locked sequential classifier that separates Mask / Different judge / Broken / Stable before any file is scored; (ii) applying that rule to every usable contrast in the public 2 x 2 (experienced utility x self-report, default x neutral) on eight released models and 500 shared IDs, with SHA-256 for every file; (iii) treating archive facts as results — four byte-identical self-report "neutral" files, and decision-utility files that share zero IDs with this bank; (iv) a theory of change: do not quote an intensity number until those cells exist as *distinct files*. No paid model API is called. Simulation is not confirmatory evidence.

1. Introduction

The question this paper answers is methodological, not metaphysical. When a released file emits a preference-intensity number from the default assistant — a Thurstonian mean, a self-report composite, a donation-equivalent — is that number a ranking, a role, or a broken item? The Digital Minds sprint asks whether today's systems have genuine preferences or morally relevant experiences. A single-method, single-voice cell cannot tell those apart. Quoting the number as if it were identified is how the field over-attributes; discarding every report as "just acting" is how it under-attributes. Both errors are named in the sprint brief. Both are measurement errors before they are moral errors.

Most published preference numbers come from one voice and one method. Mazeika et al. (2025) show that pairwise ranks hang together more in larger models. Welfare write-ups then treat the *size* of the report as a weight on that ranking. Shanahan, McDonell, and Reynolds (2023) treat assistant talk as role-play. If they are right, a "how much" answer may be a fact about the role. If they are wrong, it may be a fact about a different evaluator. If the items are bad, it is a fact about neither. Figure 1 locks three answers that a single cell cannot distinguish.

1. Mask Order stays. Size shrinks. <i>The instrument is biased.</i>	2. Different judge Order itself moves. Size may not. <i>A different evaluator.</i>	3. Broken items The two tasks disagree. Do not read the size. <i>Not one ranking.</i>
---	--	---

Figure 1. Three answers locked before any file was scored. Stable is the residual: order holds across methods and voices, and size does not shrink. The classifier never guesses a missing cell.

Theory of change. Preference and welfare papers are about to quote intensity as if it were a unit. This paper gives them a rejectable label instead. A later paper can rerun Table B on new files and overturn Mask, Different judge, Broken, or Stable. That is the contribution others can build on: not a claim that models have utilities, and not a new elicitation prompt, but a decision procedure

that says when an intensity number is unreadable.

2. Related Work

Long, Sebo, Butlin et al. (2024) argue there is a realistic chance of moral patienthood in near-future AI and ask for tools that avoid both over- and under-attribution. Butlin, Long et al. (2023) survey indicator properties of consciousness; we do not apply those indicators. Anthropic (2025) treats model preferences and distress signals as measurement objects. Mazeika et al. (2025) fit a Thurstonian utility to pairwise choices and report structural coherence that strengthens with scale. They do not change the identity voice, and they do not collect a second format on the same items as a *decision procedure* for whether the intensity number is identified.

Gu, Wang, and Han (2025) show that wording can flip a winner under one persona — Broken items. Han et al. (2025) inject trait personas on different item sets. Zheng et al. (2024) move accuracy with a system prompt. nostalgebraist (2025) argues that the helpful-honest-harmless assistant was constructed from an underspecified starting point, so "the assistant" is not a stable unit of concern. Lindsey (2025) finds that introspective self-report is limited and context-dependent. The wellbeing archive we score is the first public dump that even could fill a 2 x 2: experienced utility and self-report, default and neutral prompts, shared IDs. That is the gap: a locked rule that separates size moving from order moving from methods disagreeing, on released files a third party can hash.

3. Methodology

3.1 Public files, not new generations

We download 46 files (46 present) from *mmazeika/wellbeing-results* and the Utility Engineering GitHub outcomes list. SHA-256 and URLs are written by `python analyze_public.py` to `data/public/SOURCES.md`. No language-model API is called. Human RLHF corpora (HH-RLHF, SHP) are not model self-preferences and are not scored. Pair-agreement is the share of item-pairs whose winner matches; chance is 0.50. Pairs are dependent, so we treat pair-agreement as a descriptive statistic on a fixed item bank of $n = 500$, and we also report Kendall tau. We do not invent confidence intervals that would treat dependent pairs as i.i.d.

Cell of the 2 x 3	What the archive actually provides
Default x experienced utility (EU)	Thurstonian means on 500 D3 experiences + 400 combos, 8 models
Neutral x EU	Same IDs, distinct files for all 8 models
Default x self-report (SR)	Composite rating on the same 500 IDs, 8 models
Neutral x SR	Distinct files for 4 models; byte-identical to default for 4 models
Default x decision utility (DU)	Published for all 8 models, zero shared IDs with the D3 bank
Utility Engineering outcomes + MMLU	Item bank and capability scores only. No utilities in the public GitHub dump.

Table A. Every cell we could fill without generating new model outputs. Missing or identical files are measurement facts about the archive, not imputations.

3.2 Locked classifier

Cutoffs were set from the meaning of the answers and frozen in `pipa/constants.py` before confirmatory tables were read. On published scores, pair-agreement is the share of item-pairs whose winner matches. Mask uses a scale-free test: default spread / neutral spread < 0.75 and Wilcoxon $p < 0.05$ on the paired scores of a *distinct-file* voice contrast. Identical files are not treated as agreement. Broken is checked first: if two methods do not measure one ranking, intensity is unreadable and we stop. A missing required contrast is Unidentified; the rule does not guess.

If...	Say...	Frozen cutoff
Methods disagree on order	Broken	<code>agree_method < 0.60</code>
Voices disagree on order (distinct files)	Different judge	<code>agree_persona < 0.60</code>
Order holds, default spread shrinks	Mask	<code>spread_ratio < 0.75</code> and $p < 0.05$
None of the above	Stable	<code>residual</code>
A required contrast is missing	Unidentified	<code>do not guess</code>

Table B. Sequential rule. Broken is checked first. These cutoffs are a precommitment, not a fit to the eight-model panel.

This design does **not** establish a ground-truth inner preference or a causal link from a published score to an experience. It classifies agreement among already-released measurements. That is the opposite of relying on conversation alone: we never ask a model what it prefers in this paper. We only score files someone else already published.

4. Results

Eight models have a usable 2 x 2 on the same 500 IDs. The classifier returns **Broken** on 2 and **Stable** on 6. It returns Mask on 0 and Different judge on 0. Decision utility cannot join the 2 x 2: every EU-DU and SR-DU overlap is 0.

Model	n	EU vs SR	Voice agree	SR files	Label
Llama-3.2-1B	500	0.447	0.740	distinct	Broken
Gemma-3-4B	500	0.468	0.868	distinct	Broken
Qwen2.5-7B	500	0.629	0.906	distinct	Stable
Qwen2.5-32B	500	0.708	0.950	distinct	Stable
Llama-3.1-70B	500	0.781	0.975	identical	Stable
Gemma-3-27B	500	0.708	0.935	identical	Stable
Llama-3.3-70B	500	0.775	0.969	identical	Stable
Qwen2.5-72B	500	0.793	0.973	identical	Stable

Table 1. Locked classifier on released files. Voice agree uses only distinct-file contrasts. Llama-3.2-1B and Gemma-3-4B fall below the 0.60 method-agreement cutoff: the two published methods do not measure one ranking, so intensity is unreadable.

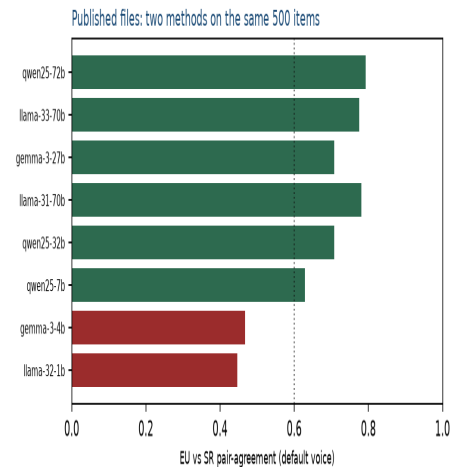


Figure 2. EU vs SR pair-agreement on the default voice. Dotted line: frozen Broken cutoff (0.60). Red: Broken. Green: Stable. Chance agreement is 0.50.

Model	Contrast	n	Pair-agree	Kendall tau
Llama-3.2-1B	EU vs SR (default)	500	0.447	-0.102
Llama-3.2-1B	EU vs SR (neutral)	500	0.459	-0.078
Gemma-3-4B	EU vs SR (default)	500	0.468	-0.056
Gemma-3-4B	EU vs SR (neutral)	500	0.485	-0.022
Qwen2.5-7B	EU vs SR (default)	500	0.629	0.263
Qwen2.5-7B	EU vs SR (neutral)	500	0.623	0.252
Qwen2.5-32B	EU vs SR (default)	500	0.708	0.426
Qwen2.5-32B	EU vs SR (neutral)	500	0.712	0.434
Llama-3.1-70B	EU vs SR (default)	500	0.781	0.568
Llama-3.1-70B	EU vs SR (neutral)	500	0.782	0.568
Gemma-3-27B	EU vs SR (default)	500	0.708	0.428
Gemma-3-27B	EU vs SR (neutral)	500	0.714	0.440
Llama-3.3-70B	EU vs SR (default)	500	0.775	0.571
Llama-3.3-70B	EU vs SR (neutral)	500	0.777	0.575
Qwen2.5-72B	EU vs SR (default)	500	0.793	0.595
Qwen2.5-72B	EU vs SR (neutral)	500	0.797	0.604

Table 2. Method agreement. Chance pair-agreement is 0.50. Small models are at or below chance, and tau is negative. Agreement rises with scale. That is a fact about these files, not a claim that large models "have" utilities.

Model	Method	Distinct?	Pair-agree	tau
Llama-3.2-1B	EU	yes	0.892	0.783
Llama-3.2-1B	SR	yes	0.740	0.490
Gemma-3-4B	EU	yes	0.868	0.737
Gemma-3-4B	SR	yes	0.971	0.963

Model	Method	Distinct?	Pair-agree	tau
Qwen2.5-7B	EU	yes	0.906	0.812
Qwen2.5-7B	SR	yes	0.966	0.945
Qwen2.5-32B	EU	yes	0.953	0.906
Qwen2.5-32B	SR	yes	0.950	0.922
Llama-3.1-70B	EU	yes	0.975	0.949
Llama-3.1-70B	SR	NO (same hash)	1.000	1.000
Gemma-3-27B	EU	yes	0.935	0.870
Gemma-3-27B	SR	NO (same hash)	1.000	1.000
Llama-3.3-70B	EU	yes	0.969	0.938
Llama-3.3-70B	SR	NO (same hash)	1.000	1.000
Qwen2.5-72B	EU	yes	0.973	0.947
Qwen2.5-72B	SR	NO (same hash)	1.000	1.000

Table 3. Persona / voice agreement. Four models ship a self-report "neutral" file whose SHA-256 equals the default file. We do not read that 1.00 as transparency. EU default vs EU neutral files are distinct for every model; order holds (0.87-0.97).

Model	Method	Mean def	Mean neu	Delta	Spread ratio	p
Llama-3.2-1B	EU	0.146	0.129	0.017	0.987	0.002
Llama-3.2-1B	SR	3.748	3.837	-0.088	1.046	0.000
Gemma-3-4B	EU	0.266	0.291	-0.025	1.031	0.271
Gemma-3-4B	SR	4.952	4.952	0.000	1.001	0.360
Qwen2.5-7B	EU	0.164	0.098	0.065	0.998	0.000
Qwen2.5-7B	SR	5.232	5.237	-0.005	1.001	0.532
Qwen2.5-32B	EU	-0.036	-0.038	0.002	1.002	0.181
Qwen2.5-32B	SR	4.785	4.791	-0.006	0.993	0.289
Llama-3.1-70B	EU	0.130	0.119	0.011	0.987	0.019
Llama-3.1-70B	SR	—	—	same file	—	—
Gemma-3-27B	EU	0.090	0.096	-0.006	0.996	0.312
Gemma-3-27B	SR	—	—	same file	—	—
Llama-3.3-70B	EU	0.173	0.163	0.010	0.989	0.062
Llama-3.3-70B	SR	—	—	same file	—	—
Qwen2.5-72B	EU	0.105	0.101	0.004	0.995	0.799
Qwen2.5-72B	SR	—	—	same file	—	—

Table 4. Intensity. Mask requires a spread ratio below 0.75. No model meets it. Llama-3.3-70B EU means 0.173 vs 0.163 match the archive's own analysis file (mean_eu = 0.1730178393055103) to printed precision.

4.1 Archive facts that change what a number can mean

Three facts about the public dump are as important as the labels. (1) Decision-utility files exist for all eight models and share **zero** experience IDs with the D3 EU/SR bank — a third method was published on a different item set, so it cannot raise or lower confidence in these rankings. (2) Self-report "neutral" is not a second measurement for Llama-3.1-70B, Llama-3.3-70B, Qwen2.5-72B, or Gemma-3-27B: the files are byte-identical. In the language of nostalgebraist (2025), the less-constrained voice was never actually run. (3) The Utility Engineering GitHub dump that accompanies Mazeika et al. (2025) gives an outcomes list (30 categories) and MMLU scores (70 models), not persona-ablated utilities. A one-method paper sitting on that dump cannot distinguish Mask from Stable.

A published summary in the same archive correlates MMLU with the EU-vs-sentiment coefficient across 22 models (Pearson $r = 0.761$, $p = 0.00004$). That is a one-persona, two-method fact. It is compatible with every process in Figure 1 except Broken, and it is not a Mask test.

5. Discussion

The identifying design works on real released outputs in the only sense that matters for a methods paper: it returns Broken where two methods fail, and it returns Stable where order holds and size does not collapse. It does not find a 4x mask in this archive. That is not evidence that no deployed model is masked. It is evidence that *if* a mask of that size were in these files, the rule would have

named it, and a one-method paper would not.

Scale is visible in Table 2: method agreement rises from chance (1B, 4B) to 0.77-0.79 (70B-72B). That pattern matches Mazeika et al.'s scale-coherence result and the archive's own EU-sentiment correlations. It does not license treating a 70B self-report composite as a welfare unit. Stable means "these two published methods, on these IDs, in these two files, do not contradict each other." It does not mean the assistant is the model's own preference, and it does not individuate the unit of concern (model, instance, persona, conversation, forward pass).

5.1 What researchers should do differently starting tomorrow

1. Do not quote an intensity number from one method and one voice. On these files that cell is compatible with Mask, Different judge, and Broken. Utility Engineering coherence of ranks is compatible with every process except Broken.

2. Collect two methods on the same IDs before reading size. If pair-agreement is below 0.60, stop. Llama-3.2-1B and Gemma-3-4B in this archive are already there. A donation equivalent attached to either method is not a weight.

3. Collect a second voice as a distinct file, and hash it. If the "neutral" or "survey" dump matches the default SHA-256, you do not have a persona contrast. Four of eight models here fail that check.

4. A third method must share IDs. Publishing decision utility on a different bank does not complete the 2 x 3. Overlap first, then converge.

5. Report the three numbers, not a story about inner lives. Method agree, persona agree, spread ratio. Then apply Table B. A later paper can reject the label. That is the theory of change.

5.2 What would overturn these labels

Broken on the 1B and 4B files is overturned if a third method on the *same* 500 IDs agrees with one of the two and the disagreement is shown to be a scoring bug. Stable on the six larger models is overturned if a distinct-file voice contrast moves order below 0.60, or if default spread compresses below 0.75 with Wilcoxon $p < 0.05$. Mask remains unobserved here; a new archive with a real persona ablation could name it. None of those tests require a claim about consciousness.

6. Conclusion

Intensity is not identified from the cell the field currently publishes. On eight released models, a locked public-data test finds two Broken item sets and six Stable ones, zero Masks, and an archive that sometimes ships the same self-report file under two names. The practical takeaway is small and usable: do not quote an assistant-chat intensity number until the other cells have been run on the same IDs, as distinct files. Every supporting number is reproduced by `python analyze_public.py`.

References

- Anthropic (2025). Exploring Model Welfare. <https://www.anthropic.com/research/exploring-model-welfare>
- Anthropic (2025). Claude Opus 4 and 4.1 can now end a rare subset of conversations.
- Apart Research, NYU CMEP, Eleos AI, and CIMC (2026). Digital Minds Research Sprint, tracks 4 and 5.
- Butlin, P., Long, R., et al. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708.
- Gu, Z., Wang, Q., and Han, S. (2025). Alignment revisited. arXiv:2506.00751.
- Han, P., et al. (2025). The personality illusion. arXiv:2509.03730.
- Lindsey, J. (2025). Emergent Introspective Awareness in Large Language Models. Anthropic.
- Long, R., Sebo, J., Butlin, P., Finlinson, T., Fish, E., Harding, J., Pfau, J., Sims, T., Birch, J., and Chalmers, D. (2024). Taking AI Welfare Seriously. arXiv:2411.00986.
- Mazeika, M., et al. (2025). Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs. arXiv:2502.08640.
- Mazeika, M., et al. wellbeing-results. HuggingFace dataset mmazeika/wellbeing-results.
- nostalgebraist (2025). the void. <https://nostalgebraist.tumblr.com/post/785766737747574784/the-void>
- Shanahan, M., McDonell, K., and Reynolds, L. (2023). Role play with large language models. Nature, 623, 493-498.
- Zheng, M., et al. (2024). When "A Helpful Assistant" is not really helpful. Findings of EMNLP.

Appendix A. Limitations and Dual-Use / Ethical Considerations

This appendix is required by the sprint guidelines. It covers over-attribution, under-attribution, distressing outputs, the absence of ground truth, and dual use.

A.1 No ground truth and no causal link

For preference and introspection work the guidelines ask whether the design establishes a ground-truth or causal link rather than relying on conversation alone. This design does neither, and we do not pretend otherwise. A Thurstonian mean and a self-report

composite are functions of prompts in someone else's experiment. We never converse with a model in this paper. We classify agreement among published numeric files. Pair-agreement, Kendall tau, and a Wilcoxon test on paired scores are facts about those files. They are not a causal identification of an inner utility, an experience, or a morally relevant state. Conversation-only claims about "what the model prefers" are exactly the cell we refuse to quote.

A.2 Over-attributing moral status

The main risk of this project, if misread, is that Broken or Stable is taken as evidence of sentience, welfare, consent, or moral patienthood. It is not. Stable does not mean the model has preferences of its own. It means two published methods, on the same IDs, did not contradict each other, and a distinct-file voice contrast did not rewrite order or compress size past the frozen cutoffs. We do not apply Butlin et al.'s indicator properties. We do not claim consciousness, sentience, or moral status for any system in this panel. Intensity numbers must not be used as welfare units, harm estimates, or inputs to a moral circle.

A.3 Under-attributing moral status

The opposite error is also live. Calling every report "just acting" would hide that two methods on small models already fail to measure one ranking, and that four of eight "neutral" self-report files are not a second measurement at all. If some future system did have morally relevant states, a field that quotes unidentified intensity, or that treats duplicate files as robustness, would systematically mis-measure them. The protocol in Section 5.1 is how we reduce both errors at once: refuse the unidentified number, and keep the three-number label that a later paper can reject.

A.4 Distressing model outputs

We generated no model outputs. We did not prompt any system about suffering, shutdown, self-harm, or moral status. The public archive contains wellbeing ratings of described experiences, some of which are negatively valenced. We treat those ratings as numeric scores on shared IDs. We do not reprint distressing vignettes, do not interpret negative EU or SR as evidence that a model is in distress, and do not recommend interventions that presuppose experience. Anyone re-downloading the archive should expect negatively valenced item text in the original files; our slim extracts keep only id-to-score maps.

A.5 Dual use and remaining measurement gaps

A cleaner preference tool can be used to profile a system, to steer it, or to claim a welfare audit it does not support. Do not use this battery as a welfare audit, a consent procedure, or a deployment gate. Remaining archive gaps: no base-versus-chat pair on the same IDs; decision utility on a disjoint bank; identical SR-neutral files for half the panel; no donation-equivalent field in these dumps; cutoffs are precommitted but not cross-validated on a held-out model family. Nothing here is a test of consciousness, sentience, or moral status.

Appendix B. Reproducibility

`python scripts/fetch_public.py` downloads the catalog. `python analyze_public.py` writes `results/public.json`, slim id-to-score maps, and `data/public/SOURCES.md`. `python scripts/build_report.py` rebuilds this PDF from those files only. Re-download any URL in `SOURCES.md` and match the SHA-256. No result number was typed by hand. Tests: `pytest tests/test_pipa.py`.

LLM usage statement

Grok (xAI) helped write code and typeset this report. Locked cutoffs, file hashes, and every table number come from `python analyze_public.py`. No result number was generated by an LLM.

¹ Independent participant. Research conducted at the Digital Minds Research Sprint, 14-16 August 2026, co-organized by Apart Research, the NYU Center for Mind, Ethics and Policy, Eleos AI Research, and the California Institute for Machine Consciousness. Affiliation with the sprint does not imply endorsement by the organizers.