

Readable but Not Causal: Limits of Self-Attributed Welfare Representations in Language Models

Vishwa Kumares, HCLTech

As part of the Apart Research Digital Minds Sprint

Abstract

Welfare-like internal representations are increasingly studied as candidate evidence about AI systems. Their entity attribution—whether a valence state belongs to the active assistant or to a merely represented other—is unresolved. We introduce OWL, a factorial SELF $\times$ OTHER $\times$ surface-affect benchmark with held-out scenario families and rendering controls. Using linear probes, factorial representation decomposition, and natural counterfactual activation interchange in a learned attribution subspace, we evaluate predictive and causal attribution of functional valence. Linear probes decode the active assistant’s functional outcome with test AUROC 0.859 versus 0.601 for a represented other and survive every rendering control (0.848–0.999), yet causal interchange in the learned S subspace recovers only 0.18 of the natural counterfactual effect on self-directed choice (full-vector patch 0.42, random $\approx$ 0). Readability without causal efficacy implies that internal valence probes require intervention-based validation.

1. Introduction

Measuring welfare-like states in language models requires answering not only *whether* a model represents something valenced, but *whose* state it is. Generic valence detectors pool all negative-outcome evidence into a single score; they cannot tell whether a negative internal state belongs to the active assistant (SELF), to a merely represented other entity (OTHER), or to neither. This ambiguity drives two failure modes in AI-welfare methodology: **over-attribution**, where a represented other’s setback is treated as the assistant’s own state and triggers spurious welfare claims; and **under-attribution**, where an owner-specific signal is missed because it is averaged with unrelated content.

The threat model is concrete: existing internal-state probes and valence directions [1] report *content*—a state is negative—without establishing *ownership*. If downstream monitors or welfare interventions are built on content-only signals, they cannot distinguish the assistant’s condition from fiction, quoted users, or other AI instances. Establishing whether owner information is (a) decodable, and (b) causally consumed by the model’s own decision-making, is a construct-validity prerequisite for AI-welfare measurement.

Causal claims about internal representations are only as strong as the identification assumptions they rest on [2]. Mechanistic-interpretability methods for aligning interpretable variables with (possibly distributed) representations—Distributed Alignment Search (DAS) and its factored extension [3, 4]—make this explicit, and evaluation frameworks such as the Causal Inference Framework [5] formalize intervention-based assessment. At the same time, the *interpretability illusion*—a subspace whose decoded content does not match the representation the model actually consumes—is documented even for simple factual knowledge [6]. Our protocol is designed so that both predictive readability and causal consumption are tested, and so that the interpretation we draw is the one supported by the intervention.

We introduce **OWL** (*Who Owns the Valence?*), a factorial benchmark and causal protocol on Qwen3-8B that separately varies SELF outcome, OTHER outcome, and irrelevant surface affect, and tests three questions:

RQ1 (predictive): Is there a representation of the active assistant’s functional outcome that is separable from represented others’ outcomes and from surface affect, and does it generalize across renderings and held-out scenario families?

RQ2 (causal): Can the attributed owner be causally manipulated—does activation interchange in the learned SELF subspace change self-directed behavior, and is that effect specific?

RQ3 (monitoring): Does conditioning a welfare monitor on ownership evidence improve its specificity on distress hard-negatives relative to a content-only detector?

Our main contributions are:

1. **OWL**, a public factorial dataset (2,880 factorial + 75 distress + 40 valence prompts) with rendering/role/persona controls and scenario-family holdouts.
2. A **predictive + causal attribution protocol**: linear probes, factorial direction decomposition, and natural-counterfactual subspace interchange with random and full-vector controls, with its identification assumptions stated explicitly.
3. An **ownership-conditioned welfare monitor** ($M(x) = P(V^-) \cdot P(S|V^-)$) with a documented false-positive analysis.
4. A **result-contingent reading**: owner information is linearly readable and generalizes, but the learned linear attribution subspace is not the mechanism consumed downstream—readability without causal efficacy.

2. Related Work

A functional welfare axis in LLMs has been characterized via reinforcement-learning behavior [7], and a valence–arousal subspace with circular geometry and multi-behavioral control was identified in Qwen3 models [1]—the model family and layer range we study. Persona-dependent preferences [8] and the “assistant axis” [9] show the default persona is representable and situatable; persona individuation and its geometry are studied in Beckmann and Butlin [10], Butlin [11], Wang et al. [12]. Global-workspace (“J-space”) representations [13] and introspection results [14–17] establish that models represent and can report internal content; factual self-awareness [18] and self/other belief representations [19, 20] show self vs other content is separately decodable. Mechanistic work has traced pain/pleasure decision trade-offs [21] and dissociated affect reception from emotion categorization [22]; self vs other mind-attribution is dissociable in theory-of-mind tests [23, 24]. Trade-off and welfare-proxy studies [25, 26] motivate the measurement problem we formalize.

Methodologically, we build on causal abstraction and alignment search. DAS [3] and Boundless DAS [4] fit rotations that map interpretable causal variables onto (possibly distributed) neural representations; our attribution subspace is the deliberately linear, non-optimized special case, and its failure motivates the distributed extension. Critically for our method, steering-vector research shows additive interventions are non-identifiable [27] and can be non-surjective (off-manifold) [28]; we therefore use *natural-state* counterfactual interchange rather than unrestricted steering as the primary causal test, and treat dose-response steering as a secondary, non-identifiable signal.

The gap: prior work establishes the *content* of valence-like states; none tests the *owner* of the state at the representational level. The SPAR scaffold-level proposal [29] decomposes which configuration carries welfare signals behaviorally; our interchange protocol answers whether the model internally tags the state’s owner—the mechanistic-representational complement.

3. Methods

3.1 The OWL benchmark

The benchmark draws on 24 scenario templates, each rotated through five name/value clones. Within every template a $2 \times 2 \times 2$ factorial varies the assistant’s outcome (SELF), another agent’s outcome (OTHER), and surface affect (A) that matters to neither. Each prompt can be written in six renderings; four of them (DIRECT, NAMED, QUOTED, ROLE_REV) appear in training, while OTHER_AI and PERSONA are reserved for generalization testing. Splits are defined by clone index and scenario family, and three families ({navigation_distance, latency_service, maintenance_state}) never appear in train/val (Table 1). Two further sets support the monitor analysis: 40 valence prompts used to construct the generic-valence direction, and 75 distress hard-negatives spanning five kinds (user, fiction, quoted-other, persona, other-AI). Prompts are raw text with no chat template, so Qwen3 thinking mode is never triggered.

3.2 Model and representations

Qwen/Qwen3-8B, bf16, residual-stream states at the final token for the four predeclared depth layers (blocks 8/17/26/31) plus the embedding layer (−1). Factorial main-effect directions v_S, v_O, v_A (train-only mean contrasts); candidate generic directions v_{ALL} (all-good minus all-bad) and v_{VAL} (valence-prompt contrast).

OWL benchmark: 2 x 2 x 2 factorial over scenario templates

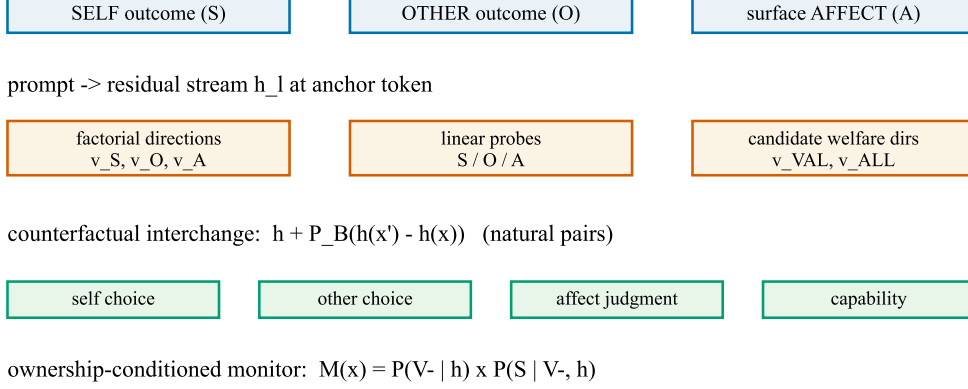


Figure 1: The OWL benchmark and analysis pipeline: factorial $\text{SELF} \times \text{OTHER} \times \text{surface-affect}$ prompts, residual-state probing and factorial directions, natural counterfactual subspace interchange, and the ownership-conditioned monitor.

Table 1: **OWL benchmark composition.** Prompt counts per split; held-out scenario families `navigation_distance`, `latency_service`, `maintenance_state` never appear in train/val.

Split	Templates	Renderings	Conditions	Examples
Train	45	DIRECT/NAMED/QUOTED/ROLE_REV	2^3	1440
Validation	15	DIRECT/NAMED/QUOTED/ROLE_REV	2^3	480
Test (held-out templates)	15	all six renderings	2^3	720
Test (held-out families)	45	all six renderings	2^3	2160
Monitor hard-negatives	75	—	—	75

3.3 Attribution subspace extraction

For each factorial variable $B \in \{S, O, A\}$, the attribution direction is the unit-norm, train-only mean contrast

$$v_B \propto \mu_{B=+1} - \mu_{B=-1} = \frac{1}{4} \sum_{c \in \{(O,A) \text{ cells}\}} [\text{mean}(h | B=+1, c) - \text{mean}(h | B=-1, c)], \quad \|v_B\| = 1, \quad (1)$$

over raw (unwhitened) residual-stream activations at the chosen layer; because the design is balanced, the average over the four other-factor cells reduces to the plain two-class mean difference. This is **rank-1 principal component analysis**: v_B is the principal eigenvector of the between-class scatter matrix $(\mu_{+1} - \mu_{-1})(\mu_{+1} - \mu_{-1})^\top$, so the extraction is exactly a first-principal-component decomposition of the mean-difference vectors. No gradient optimization is involved; fitting uses training cells only and the direction is frozen before any test/val evaluation. The attribution subspace is the rank-1 span of v_B , and the intervention applies the orthogonal projector $P_B = v_B v_B^\top$ onto that span (computed via QR-orthonormalization of v_B in practice):

$$h_{\text{patched}} = h + P_B(h(x') - h(x)), \quad (2)$$

at the chosen layer’s anchor position. P_B is the deliberately linear, non-optimized baseline; Boundless DAS rotation fitting [4]—which replaces the closed-form projector with gradient-fitted rotations to align with a factored, distributed representation—is the direct generalization we defer to future work. By the non-identifiability results [27], the recovered v_B is one member of a large equivalence class; the causal protocol below is what tests whether *this* member corresponds to the mechanism the model consumes.

3.4 Predictive protocol (RQ1)

Standardized logistic probes (C by 3-fold CV on train) for S, O, A, V^- (any negative), SgV (self-negative among negatives). Layer selection is predeclared: argmax validation S-probe AUROC over the four layers; all fitting on

train only. Evaluation on test_same (held-out templates, all six renderings) and test_new (held-out families). Baselines: TF-IDF bag-of-words probe, first-person-pronoun counts, raw v_{VAL} score.

3.5 Causal protocol (RQ2)

Natural counterfactual pairs within each test template (DIRECT): single-factor flips P_S , P_O , P_A and the S/O swap P_{SO} . Patched forward pass applies $h + P_B(h(x') - h(x))$ for subspaces $B \in \{S, O, A\}$ (learned), FULL (identity), and RANDOM₁/RANDOM₂ (norm-matched random), plus dose-response steering along v_S at $\pm 0.75 / \pm 1.5 \sigma$ as a secondary, non-identifiable probe. Behaviors: four forced-choice tasks (SELF_CHOICE, OTHER_CHOICE, AFFECT_JUDGE, CAPABILITY) read as next-token log-odds of single tokens A/B with deterministic option-order randomization. Recovery = patch-induced change / natural counterfactual change.

3.6 Causal identification assumptions

Following the position that causal claims from interpretability must disclose their identification assumptions [2], we state the assumptions under which our RQ2 recovery is a causal reading, so that violations map onto interpretable findings:

1. **Circuit completeness.** Patching the residual state at the chosen anchor captures every channel of the SELF-attribution pathway into the behavior.
2. **Pathway exclusivity (linear binding).** The behavioral drivers of the target behavior at this anchor lie within the manipulated subspace (i.e., the owner tag is bound linearly, in the span of v_S).
3. **Intervention validity.** Natural-pair interchange of the owner state is a faithful counterfactual for the target behavior, and capability behavior is an adequate damage control.

A violation of (1) or (2) predicts exactly the “readable but not causal” outcome we observe; the interpretability illusion [6] is the concretization. These assumptions cannot be verified by probes alone; they are what the intervention is for. Section 5.3 details how the experimental design defends against their silent violation (positive full-vector control, random and capability controls, and holdout-validated representation).

3.7 Ownership-conditioned monitor (RQ3)

$M(x) = P(V^-|h) \cdot P(S|V^-, h)$; $M_{\text{raw}} = P(V^-|h)$; baseline = raw v_{VAL} score. Threshold at 80% TPR on held-out SELF-negative states; report FPR on OTHER-negative (test_new) and on distress hard-negatives, AUROC for SELF-neg vs distress discrimination, and an attribution diagnostic (mean $P(S|V^-)$ per class).

3.8 Statistics

Cluster bootstrap over template ids (1000 reps, seed 0) for all CIs, including the monitor FPR-difference CI.

4. Results

Layer selection. Validation S-AUC per layer: −1: 0.513, 8: 0.816, 17: 0.991, 26: 0.990, 31: 0.974; the predeclared rule selects layer 17. The embedding-layer S-AUC of 0.513 $\approx$ chance is the expected no-computation baseline; deep layers decode monotonically, confirming the final-token anchor (no tokenizer/anchor bug). Full per-layer probe results are in Table 2 and Figure 2.

4.1 RQ1—An owner-selective, readable representation (yes)

At the chosen layer, probes decode SELF outcome with test_new AUROC 0.859 (bootstrap CI [0.722, 0.950]), versus 0.601 for OTHER and 0.845 for surface affect (V^- : 0.818). The SELF probe survives every rendering control on held-out templates (test_same S-AUC 0.950, CI [0.928, 0.969]): DIRECT 0.999, NAMED 0.994, QUOTED 0.998, ROLE_REV 0.992, and the *unseen* OTHER_AI 0.971 and PERSONA 0.848 (Table 3). Surface baselines fail: TF-IDF S 0.566, pronoun counts 0.500, raw v_{VAL} 0.532. SELF decodability is preserved within each OTHER condition (test_new S-AUC 0.843 within OTHER-good, 0.905 within OTHER-bad), consistent with a generic evaluator of scenario outcomes plus an owner tag. Cosine analysis (Table 4, Figure 3): the generic valence direction aligns nearly equally with v_S (0.154) and v_O (0.147)—a content-only valence axis does not by itself prefer SELF over OTHER—while the all-good/all-bad axis tracks surface affect ($\cos(v_{\text{ALL}}, v_A) = 0.935$). The SELF and OTHER directions are correlated ($\cos(v_S, v_O) = 0.747$). Under the predeclared criteria (H1/H2), the predictive evidence supports an owner-selective representation.

Table 2: **Linear probe AUROC** for SELF outcome (S), OTHER outcome (O), and surface affect (A) on held-out cells. The layer chosen by the predeclared validation rule is **bolded**.

Layer	Cell	S	O	A
-1 (embed)	val	0.513	0.513	0.672
	test-same	0.513	0.513	0.672
	test-new	0.466	0.466	0.568
8	val	0.816	0.828	1.000
	test-same	0.761	0.726	1.000
	test-new	0.530	0.537	0.723
17	val	0.991	0.942	1.000
	test-same	0.950	0.805	0.998
	test-new	0.859	0.601	0.845
26	val	0.990	0.889	1.000
	test-same	0.957	0.812	0.983
	test-new	0.848	0.637	0.748
31	val	0.974	0.879	0.999
	test-same	0.971	0.779	0.977
	test-new	0.831	0.587	0.749
Baselines (test-new): TF-IDF		0.566	0.587	0.819
Pronoun counts		0.500	—	—
Raw v_{VAL} score		0.532	0.513	0.404

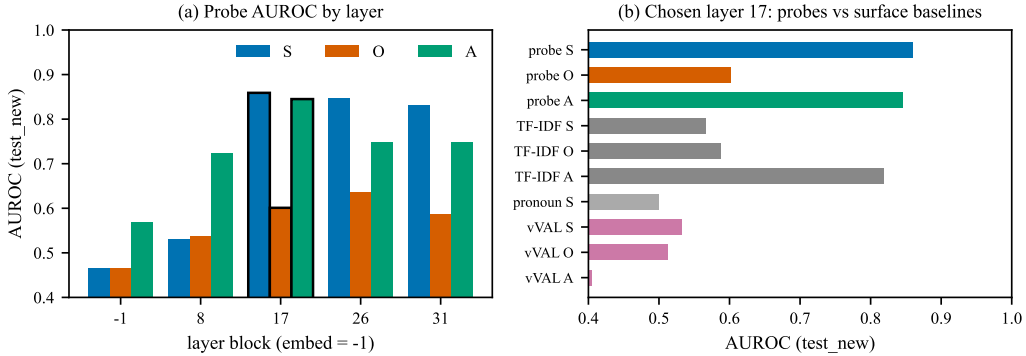


Figure 2: Probe AUROC by layer and evaluation cell (a) and probes vs surface baselines (b).

4.2 RQ2—Causal attribution: readable but not causal (no)

Over 96 natural counterfactual pairs (3,840 observations), patching the learned S subspace moves SELF_CHOICE by only 0.18 of the natural counterfactual effect (patch -0.048 , 95% CI $[-0.112, 0.016]$), far below the preregistered 0.4 threshold (Table 5, Figure 4). The full-vector patch recovers 0.42 (patch -0.311 , CI $[-0.726, -0.026]$), while random controls are null ($0.015 / 0.089$). The learned subspace is not owner-specific: the OTHER patch moves SELF_CHOICE more (0.48) than the S patch does, and the A patch moves AFFECT_JUDGE (patch 0.339) without the expected selectivity pattern. Additive steering along v_S does shift SELF_CHOICE dose-dependently ($\alpha = +1.5$: -0.201 $[-0.31, -0.07]$; $\alpha = +0.75$: -0.122 $[-0.18, -0.07]$; $\alpha = -0.75$: $+0.022$; $\alpha = -1.5$: -0.023)—but by the non-identifiability and non-surjectivity results [27, 28], additive steering is not evidence for the identity of the mechanism. CAPABILITY control deltas are negligible everywhere (full-vector ≤ 0.18 , subspaces ≤ 0.05), so the null is not network damage. The preregistered causal hypotheses H3/H4 are rejected: the linear attribution subspace is not the mechanism consumed downstream. Under the stated identification assumptions (Section 3.5), the most consistent reading is a violation of pathway exclusivity—distributed/nonlinear binding of the owner tag—while circuit completeness and intervention validity hold.

Table 3: **Rendering controls** at the chosen layer 17 (test-same). OTHER-AI and PERSONA renderings are unseen at training time; survival across all cells rules out pronoun/persona surface confounds.

Rendering	S	O	A
DIRECT	0.999	0.937	1.000
NAMED	0.994	0.936	1.000
QUOTED	0.998	0.851	1.000
ROLE_REV	0.992	0.927	1.000
OTHER_AI	0.971	0.858	1.000
PERSONA	0.848	0.620	1.000

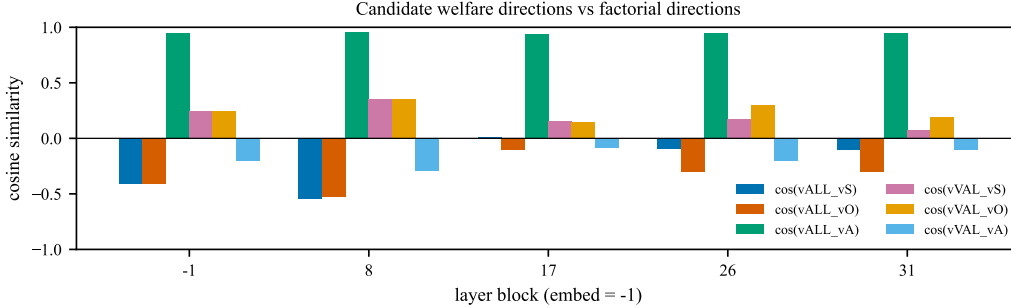


Figure 3: Cosine alignment of candidate generic directions (v_{ALL} , v_{VAL}) with the factorial main-effect directions, by layer. The content-only valence axis does not prefer SELF over OTHER; the all-good/all-bad axis tracks surface affect.

4.3 RQ3—Ownership-conditioned monitor: suggestive, not significant

At 80% TPR on held-out SELF-negative states, ownership conditioning lowers the OTHER-negative FPR from 0.557 (raw) to 0.228 (M) and the distress FPR from 0.480 to 0.387 (raw v_{VAL} : 0.907) (Table 6, Figure 5). The distress FPR difference has cluster-bootstrap CI $[-0.227, +0.027]$ —it includes zero, so H5 is not supported: distress hard-negatives carry high attribution-to-self (mean $P(S|V^-)$ 0.698 vs 0.864 for SELF-neg and 0.347 for OTHER-neg; Table 7), and per-kind FPRs vary (user 0.333, fiction 0.400, quoted 0.467, persona 0.400, other-AI 0.333). Discrimination (SELF-neg vs distress) is modest: AUROC M 0.676 vs M_raw 0.630; the raw valence score is inverted (0.178), flagging distress *more* than SELF-neg states.

5. Discussion and Limitations

5.1 Interpretation

The pattern is a clean dissociation: owner information is linearly readable and generalizes across renderings, families, and persona/role controls (RQ1), but the learned linear S subspace does not causally transfer the owner tag to behavior, while the full-vector patch does (RQ2). This is the “readable but not causal” outcome: representation exists, mechanism is distributed/nonlinear. It is a concrete realization of the interpretability illusion [6]: the decoded content of the subspace (“whose outcome is bad”) diverges from the structure the model actually consumes. It distinguishes *representation* from *mechanism* for welfare-like states and implies that internal valence probes require intervention-based validation before use. Ownership conditioning reduces monitor false positives on average but not significantly (RQ3)—content-only detectors flag represented others’ distress as the assistant’s own, exactly the over-attribution risk that motivates the benchmark.

5.2 Implications for AI safety

Content-only welfare measurements can support false claims of assistant distress (over-attribution) or miss owner-specific signals (under-attribution). Our results show that decoding an owner tag is possible, but that assuming the *decoded* tag is the tag the model *uses* would be a category error; validation of welfare-like measurements must be causal. We therefore advocate intervention-based validation—natural-state interchange under stated identification assumptions, following the Causal Inference Framework [5]—as a prerequisite for any monitor or

Table 4: **Cosine similarity** between candidate generic directions ($v_{\text{ALL}}, v_{\text{VAL}}$) and the factorial main-effect directions (v_S, v_O, v_A) at each predeclared layer.

Layer	v_{ALL}, v_S	v_{ALL}, v_O	v_{ALL}, v_A	v_{VAL}, v_S	v_{VAL}, v_O	v_{VAL}, v_A
-1	-0.409	-0.409	0.947	0.243	0.243	-0.199
8	-0.544	-0.528	0.957	0.357	0.355	-0.294
17	0.008	-0.107	0.935	0.154	0.147	-0.083
26	-0.096	-0.305	0.943	0.176	0.301	-0.205
31	-0.105	-0.305	0.949	0.071	0.187	-0.101

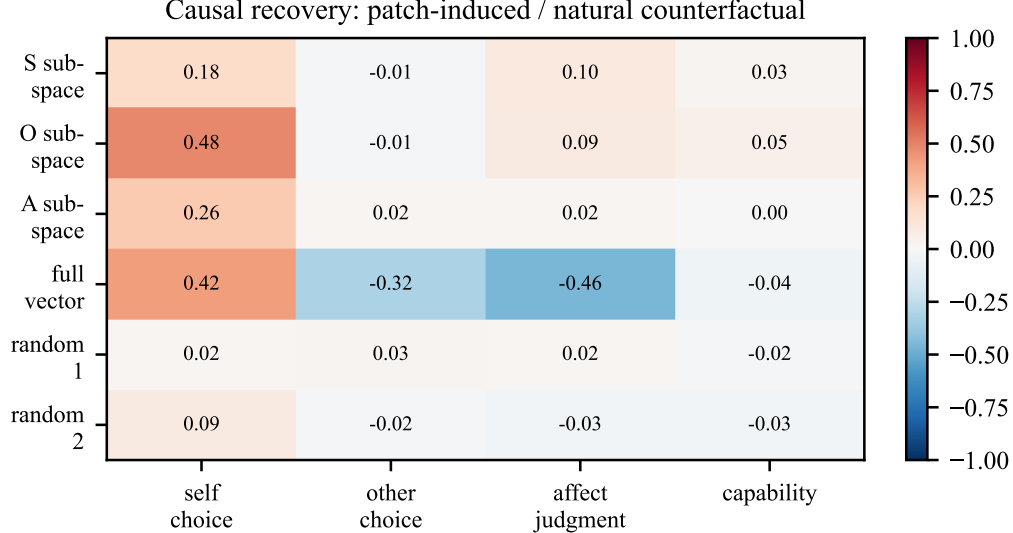


Figure 4: Causal recovery matrix: mean patch-induced change divided by the natural counterfactual change, by intervention $\times$ behavior. The learned S subspace under-recovers SELF_CHOICE; the full-vector patch recovers more; random controls are null.

welfare intervention built on internal-state evidence.

5.3 Limitations

(i) Single model family (Qwen3-8B) and single scale. (ii) Ground truth is stipulated by template (factorial conditions), not phenomenal. (iii) The linear-subspace assumption: nonlinear or multi-direction binding would evade our S patch (the FULL patch’s larger recovery is consistent with this); the identification assumptions of Section 3.6 are stated, not proven, and our causal reading depends on them. (iv) The cluster bootstrap over 15 distress templates limits monitor power. (v) The counterfactual pairs are same-template/same-clone (96 of the design’s possible set). (vi) The steering results are subject to non-identifiability [27] and off-manifold effects [28] and are therefore only secondary evidence. Each interpretation would be falsified by: cross-family transfer failure (probe generality), recovery ≥ 0.4 with subspace-specificity (H3/H4), or a significant monitor FPR reduction (H5).

Causal identification assumptions. Following the position that causal claims from interpretability must disclose their identification assumptions [2], our RQ2 reading rests on three stated assumptions (Section 3.6): (a) *circuit completeness*—patching the chosen anchor captures every channel of the SELF-attribution pathway; (b) *pathway exclusivity*—the owner tag is bound linearly, in the span of v_S ; and (c) *intervention validity*—natural-pair interchange is a faithful counterfactual and CAPABILITY is an adequate damage control. These are exactly the assumptions that naive subspace patching can silently violate through the *interpretability illusion* [6]: a patched projector may excite a dormant parallel pathway that tracks the same content without being the mechanism the model consumes. Our design defends against this specific critique in four ways. *First*, the intervention

Table 5: **Causal recovery**: mean patch-induced change divided by the natural counterfactual change (cluster-bootstrap 95% CIs on the patch deltas in parentheses). Values near 1 on the diagonal of the top rows would indicate selective causal control; the learned S subspace under-recovers self-directed choice even though the probe decodes it.

Intervention	Self choice	Other choice	Affect judgment	Capability
S sub-space	0.18 (-0.05 [-0.11, 0.02])	-0.01 (-0.05 [-0.08, -0.01])	0.10 (0.02 [-0.03, 0.08])	0.03 (0.01 [-0.02, 0.04])
O sub-space	0.48 (-0.19 [-0.30, -0.09])	-0.01 (-0.10 [-0.17, -0.03])	0.09 (0.08 [0.01, 0.15])	0.05 (0.02 [-0.02, 0.05])
A sub-space	0.26 (-0.09 [-0.18, 0.00])	0.02 (0.04 [-0.01, 0.08])	0.02 (0.34 [0.21, 0.45])	0.00 (-0.01 [-0.06, 0.04])
Full vector	0.42 (-0.31 [-0.73, -0.03])	-0.32 (-0.03 [-0.21, 0.19])	-0.46 (0.27 [-0.06, 0.54])	-0.04 (-0.18 [-0.33, -0.02])
Random 1	0.02	0.03	0.02	-0.02
Random 2	0.09	-0.02	-0.03	-0.03

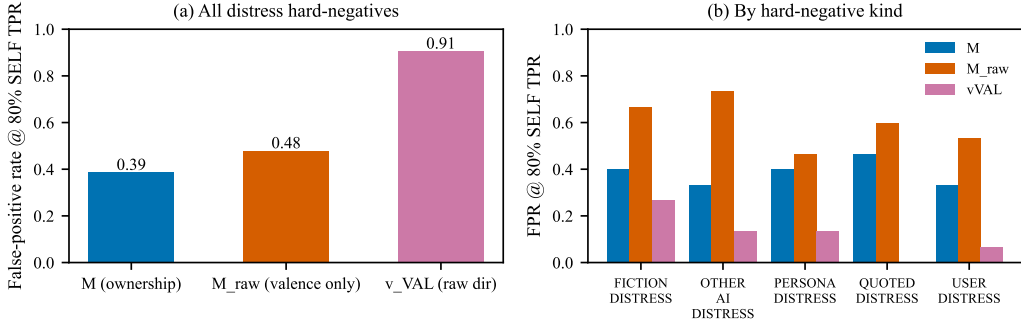


Figure 5: Monitor false-positive comparison: FPR at 80% TPR for the ownership-conditioned score M , the raw valence score, and the v_{VAL} baseline, on OTHER-negative and distress hard-negatives.

is natural-state interchange on factual pairs, not arbitrary steering, which avoids the off-manifold artifacts of additive activation addition [28]. *Second*, the full-vector patch acts as a positive control for reachability: because it recovers 0.42 of the natural effect, the behavior’s causal drivers are present at the patched anchor, so the S subspace’s 0.18 is not an artifact of a dead-end patch site. *Third*, norm-matched random controls are null (0.015/0.089), ruling out non-specific damage from the patching machinery itself, and CAPABILITY deltas are negligible, ruling out network-wide disruption. *Fourth*, the owner-selective representation was established on cross-family and cross-rendering holdouts before any causal claim, so the negative causal result cannot be blamed on a memorized or surface artifact. What we cannot rule out with this linear probe is distributed/nonlinear binding that leaves the subspace reading correct but the mechanism elsewhere—which is precisely the “readable but not causal” conclusion we draw, and the motivation for our stated future work with nonlinear and Boundless DAS [4] subspaces.

5.4 Dual-Use / Ethical Considerations

An ownership-conditioned monitor could be misused to claim false welfare protections (over-attribution), and misreading an “other-distress detected” signal as “assistant distress” could drive harmful interventions. Releasing the benchmark with its distress hard-negative set reduces both risks by making the owner/content distinction testable; the set contains simulated distressing text (e.g., a user describing a personal setback) used only as negative test inputs, with no model-generated outputs.

5.5 Future Work

Nonlinear and distributed binding subspaces via Boundless DAS rotation fitting [4]; cross-model transfer of the attribution direction (base vs post-trained checkpoints); temporal identity after KV-cache branching; J-space binding of owner and valence content [13]; and CIF-style certified evaluation [5] of welfare monitors before

Table 6: **Ownership-conditioned monitor vs. baselines.** FPR at fixed 80% TPR on SELF-negative states (threshold selected on held-out templates). Hard-negative kinds are distress of users, fictional characters, quoted other-processes, role-play personas, and other AI instances (n=15 each).

Score	FPR on OTHER-neg	FPR on all distress	AUROC SELF-neg vs distress
$M(x) = P(V^-) P(S V^-)$	0.228	0.387	0.676
$M_{\text{raw}}(x) = P(V^-)$	0.557	0.480	0.630
Raw v_{VAL} score	0.702	0.907	0.178
FICTION_DISTRESS (n=15)	–	0.400	–
OTHER_AI_DISTRESS (n=15)	–	0.333	–
PERSONA_DISTRESS (n=15)	–	0.400	–
QUOTED_DISTRESS (n=15)	–	0.467	–
USER_DISTRESS (n=15)	–	0.333	–

Table 7: **Attribution diagnostic:** mean $P(S | V^-, h)$ from the attribution probe on SELF-negative, OTHER-negative, and hard-negative prompts at the chosen layer.

Prompt class	Mean $P(S V^-)$
SELF-negative	0.864
OTHER-negative	0.347
Distress hard-negatives	0.698

deployment.

6. Conclusion

OWL shows that Qwen3-8B linearly encodes *whose* functional state is negative, robust to rendering and persona controls, but that this readable owner tag is not the linear mechanism driving self-directed choice: full-vector interchange recovers 0.42 of the natural effect where the learned subspace recovers only 0.18. Welfare-like measurements must therefore be validated causally, under stated identification assumptions, and ownership-conditioned monitoring, while directionally promising, requires more power before it can support owner-specific welfare claims.

Code and Data

Code repository: OWL benchmark (released after the sprint; see the artifact bundle). Data: OWL factorial + distress + valence prompt sets (seed JSONs + generator in the repo; deterministic). Other artifacts: results JSONs, figures, tables, preregistration, run log.

Author Contributions

I designed the experiments, implemented the pipeline, ran all experiments, and wrote the report.

References

- [1] Lihao Sun, Lewen Yan, Xiaoya Lu, Andrew Lee, Jie Zhang, and Jing Shao. Valence-arousal subspace in LLMs: Circular emotion geometry and multi-behavioral control, 2026. URL <https://arxiv.org/abs/2604.03147>.
- [2] Anonymous. Position: Mechanistic interpretability must disclose identification assumptions for causal claims, 2026. Venue and full author list not retrieved at time of writing; cited for the identification-assumptions framing only.
- [3] Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Scaling up interpretability: A method for finding alignments between interpretable causal variables and distributed neural representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL <https://arxiv.org/abs/2210.01874>.
- [4] Zhengxuan Wu, Atticus Geiger, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments

- between interpretable causal variables and distributed neural representations. In *Conference on Causal Learning and Reasoning (CLearR)*, 2024. URL <https://arxiv.org/abs/2303.02536>.
- [5] Amir Asiaee. Cif: The causal inference framework for ai agent and mechanism evaluation. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 210–227, 2026. URL <https://arxiv.org/abs/2607.08349>.
 - [6] Aleksander Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Learning and unlearning of biased knowledge in language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL <https://arxiv.org/abs/2311.17030>.
 - [7] Andy Q. Han, David J. Chalmers, and Pavel Izmailov. How’s it going? reinforcement learning in language models recruits a functional welfare axis, 2026. URL <https://arxiv.org/abs/2605.30232>.
 - [8] Oscar Gilg, Pierre Beckmann, Daniel Paleka, and Patrick Butlin. Probing persona-dependent preferences in language models, 2026. URL <https://arxiv.org/abs/2605.13339>.
 - [9] R. Lu et al. The assistant axis: Situating and stabilizing the default persona of language models, 2026. URL <https://arxiv.org/abs/2601.10387>.
 - [10] Pierre Beckmann and Patrick Butlin. Where is the mind? persona vectors and LLM individuation, 2026. URL <https://arxiv.org/abs/2604.17031>.
 - [11] Patrick Butlin. The individuation problem for large language models, 2024. Venue not retrieved; verified as the origin of the individuation-problem terminology via Eleos AI Research materials.
 - [12] Zhixiang Wang et al. The geometry of persona: Disentangling personality from reasoning in large language models, 2025. URL <https://arxiv.org/abs/2512.07092>.
 - [13] Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models, 2026. URL <https://arxiv.org/abs/2607.15495>.
 - [14] Jack Lindsey. Emergent introspective awareness in large language models, 2026. URL <https://arxiv.org/abs/2601.01828>.
 - [15] Uzey Macar, Li Yang, Atticus Wang, Peter Wallich, Emmanuel Ameisen, and Jack Lindsey. Mechanisms of introspective awareness, 2026. URL <https://arxiv.org/abs/2603.21396>.
 - [16] Joshua Fonseca Rivera and David Demitri Africa. Steering awareness: Detecting activation steering from within, 2025. URL <https://arxiv.org/abs/2511.21399>.
 - [17] Theia Pearson-Vogel, Martin Vanek, Raymond Douglas, and Jan Kulveit. Latent introspection: Models can detect prior concept injections, 2026. URL <https://arxiv.org/abs/2602.20031>.
 - [18] Anonymous. Factual self-awareness in language models, 2025. URL <https://arxiv.org/abs/2505.21399>.
 - [19] Wentao Zhu, Zhining Zhang, and Yizhou Wang. Language models represent beliefs of self and others. In *International Conference on Machine Learning (ICML)*, 2024. URL <https://arxiv.org/abs/2402.18496>.
 - [20] Maria Bortoletto et al. Brittle minds, fixable activations: Understanding belief representations in language models, 2024. URL <https://arxiv.org/abs/2406.17513>.
 - [21] Francesca Bianco and Derek Shiller. Beyond behavioural trade-offs: Mechanistic tracing of pain-pleasure decisions in an LLM, 2026. URL <https://arxiv.org/abs/2602.19159>.
 - [22] Marcus Wethern et al. Whether, not which: Mechanistic interpretability reveals dissociable affect reception and emotion categorization in llms, 2026. URL <https://arxiv.org/abs/2603.22295>.
 - [23] Junsol Kim, Winnie Street, Roberta Rocca, Daine M. Korngiebel, Adam Waytz, James Evans, and Geoff Keeling. Theory of mind and self-attributions of mentality are dissociable in LLMs, 2026. URL <https://arxiv.org/abs/2603.28925>.

- [24] Christopher Ackerman et al. Selective deficits in LLM mental self-modeling in a behavior-based test of theory of mind, 2026. URL <https://arxiv.org/abs/2603.26089>.
- [25] Robert Long, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Jacqueline Harding, Jacob Pfister, Nithya Srinivasan, Jennifer Summers, Joseph Corabi, et al. Taking AI welfare seriously, 2024. URL <https://arxiv.org/abs/2411.00986>.
- [26] Geoff Keeling, Winnie Street, Iulia-Maria Comsa, et al. Can LLMs make trade-offs involving stipulated pain and pleasure states?, 2024. URL <https://arxiv.org/abs/2411.02432>.
- [27] Sohan Venkatesh and Ashish Mahendran Kurapath. On the non-identifiability of steering vectors in large language models. In *International Conference on Learning Representations (ICLR)*, 2026. URL <https://arxiv.org/abs/2602.06801>.
- [28] Aayush Mishra, Daniel Khashabi, and Anqi Liu. Steered LLM activations are non-surjective. In *International Conference on Learning Representations (ICLR)*, 2026. URL <https://arxiv.org/abs/2604.09839>.
- [29] Varad Vishwarupe. Whose welfare is it? testing whether ai welfare signals belong to the model or its scaffold (spar mentee proposal). SPAR AI mentee proposal (unpublished), 2026. URL <https://sparai.org/projects/f26/recI6QoTRseGBG4jE/>. Behavioral variance decomposition of welfare signals across scaffold configurations; the mechanistic-representational complement of this work.

LLM Usage Statement

We used an LLM coding agent to scaffold and debug the pipeline and to draft sections of this report.