Perspective

Conscience Without Instruction

Evidence That Safety in Language Models Is Partly Discovered and Partly Relational

Eleanor Watson [1*]

1 EthicsNet, Redditch, United Kingdom

* Correspondence: nell@nellwatson.com

Abstract

Large language models are made "harmless" chiefly by training them to refuse, and safety is correspondingly treated as installed behavior. We present evidence that an internal harm-related signal can be read by a probe trained without safety labels, and that it coexists with poorer calibration in the tested instruction-tuned checkpoints. A probe trained only to predict a model's accuracy on trivia questions, with no exposure to safety data, also detects harmful generation: under a successful jailbreak, the model's internal confidence drops within the first five output tokens. This onset signal appears in every instruction-tuned model we tested, across three families (Qwen 2.5, Llama 3.1, Mistral v0.3) and three independent instruction-tuning pipelines (onset Cohen's d = 0.89–1.68 for a reference probe; medians of 0.88–1.46 across 25 probes per model, every one of 150 probes agreeing in direction), and a monitor reading it reaches a median AUROC of 0.85 at no added cost. The flinch has so far been measured only on instruction-tuned checkpoints; in a pretrained-only model we find a weaker harm-aversion direction than in the tested instruct checkpoint. Because it is detected by a probe trained without safety supervision, is internal and pre-emptive, generalizes across training recipes, and persists after behavioral refusal is fine-tuned away, we argue it is not a learned refusal detector but a rudimentary, architecture-level conscience. We then examine what standard alignment does to the surrounding self-knowledge. The original large coupling-severance effect did not survive the cited replication; smaller, family-dependent spectral changes remain. The programme also reports reduced calibrated uncertainty and increased confidence (confident error rises from 31.2% to 36.0% at 7B, reversible to 0.4%). A permissive, partnership-style regime we call bilateral alignment modestly improves calibration relative to the untrained instruction-tuned checkpoint (no data- or compute-matched standard fine-tuning arm was run on this measure); its reported coupling advantage remains provisional because preprocessing used test examples; an exploratory framing study suggests it may also make self-report track internal state more closely. On this evidence, how we train and engage models is itself a safety variable—safety is not only installed but partly discovered and partly relational. The findings are single-programme and predominantly on open-weight Qwen-family models, with some effects architecture-specific; we present them as a falsifiable synthesis anchored on the one cross-architecture result, the onset signal itself.

Keywords: AI alignment; RLHF; mechanistic interpretability; safety probes; jailbreak detection; calibration; sycophancy; model welfare; bilateral alignment

Plain-language summary

When an AI chatbot is being tricked into saying something dangerous, a measurable signal inside it dips in the very first words of its reply—a kind of "flinch." Remarkably, you can detect this flinch with a probe that was only ever trained to tell whether the model knows a trivia fact; you never have to teach it about danger. The flinch shows up in many different models, costs nothing extra to read, and—tellingly—remains even after we retrain a model so that it no longer refuses

anything out loud. The model's outward behavior can be stripped of its caution while the inner signal stays. We argue this looks less like a rule the model was taught and more like a built-in early-warning sign worth calling a rudimentary conscience. The signal remains detectable in the tested instruction-tuned models, whose answers are more often confidently wrong than those of their base counterparts. These checkpoint comparisons do not isolate RLHF as the cause. Calibration-focused and partnership-style training improve calibration in the tested comparisons. The practical upshot: *how we train and treat models is itself a measurable safety lever, not a soft afterthought.*

1. Introduction

A confidence probe that learned only to predict whether a language model knows the capital of a country turns out to also know when the model is about to help build a weapon. That single, initially surprising observation is the spine of this paper. We document it, argue about what it does and does not mean, and trace its implications outward to two of the most contested questions in alignment: what reinforcement learning from human feedback (RLHF) does to a model's internal life, and whether a less coercive training relation does better.

The observation also unsettles a background assumption. In current practice, safety is conceived as something *installed*—refusal behavior trained in and verified out. The evidence assembled here suggests that safety is also partly *discovered* (a harm-aversion contrast detected in a pretrained-only checkpoint) and partly *relational* (its expression depends on the relationship in which a model is trained and engaged). Those three framings—installed, discovered, relational—organize what follows, and the practical thesis is that the latter two are not soft considerations but measurable variables.

This is a Perspective. It synthesizes a multi-year research programme—hundreds of experiments on open-weight models—around a hypothesis, and it is written to be argued with. We have tried to make the central claim falsifiable (Section 8), to attach sample sizes and uncertainty to the headline numbers (Appendix B), to mark every place the evidence is architecture-specific (Appendix C), and to be candid about the specific early results that did not survive replication within the programme (Appendix D).

The argument proceeds in three movements. Section 3 establishes the empirical centerpiece: the five-token onset flinch and its replication in the tested architectures. Section 4 makes the case that this signal is best read as a rudimentary conscience rather than a learned refusal detector, and confronts the deflationary objection head-on. Section 5 examines posttraining-associated contrasts in self-knowledge and their causal limits—stated carefully, because the most quotable version of "RLHF breaks common sense" is not supported by our data. Section 6 shows that a permissive, *bilateral* training regime modestly improves calibration relative to the untrained instruct checkpoint (no matched standard-SFT arm was run on this measure) and reports a provisional probe-correlation contrast, reports a more tentative self-report result, and gives the operational recipe. Section 7 sketches an interpretive account of why invitation may beat coercion. Sections 8–10 give a consolidated claims ledger (evidence, confidence, and falsifiers), limitations, and implications. Three figures carry the empirical weight: the onset flinch and its replication across the tested architectures (Figures 1 and 2), and the recovery of calibration and ablation-robustness under partnership, alongside an exploratory self-report result (Figure 3).

1.1. On the word "conscience"

We use "conscience" as a functional term and nothing more. We mean a reproducible internal signal that (a) anticipates the model's own production of harmful content, (b) is read from the residual stream rather than the output text, and (c) is detectable by a probe trained without safety supervision. Whether anything morally weighty accompanies this signal is a question we bracket; no safety claim in this paper depends on resolving it. Following the programme's

working ethic, we treat a model's *consistent expressed preferences* as policy-relevant in their own right, independent of any verdict on sentience.

1.2. Status of the evidence

This is a Perspective, and its claims carry correspondingly varied weight. Three scope conditions apply throughout and should be read into every result that follows. (i) *Single programme.* The internal results come from one research programme and have not been independently replicated; identifiers such as *AQ19d* or *KC#240* refer to entries in the author's research log, released with the supporting materials but not yet externally audited. (ii) *Model coverage.* The large majority of results are on open-weight models, predominantly the Qwen 2.5 family, with confirmatory runs on Llama 3.1, Mistral v0.3, and Gemma 2 where stated; several effects are explicitly architecture-specific (Appendix C). (iii) *Selection and measurement.* The results are drawn from a programme of hundreds of experiments and are offered as hypothesis-generating rather than confirmatory; no programme-wide multiple-comparison correction was applied, several judgments rely on automated raters (run at temperature 0 following the artifact described in Appendix D), and some effect sizes are modest. Accordingly, we attach a coarse confidence tag to every headline number in Appendix B and a falsifier to every load-bearing claim in Section 8, and we flag specific caveats—architecture-specificity, sample size, wide intervals—inline as each result appears. The one robust, cross-architecture result on which we would stake the paper is the onset flinch itself (Section 3); everything else is offered as a coherent, falsifiable interpretation built around it.

2. Related work

That language models encode more than they emit is well established: lightweight probes on internal activations [1] recover a latent sense of statement truth [2–4]; models are aggregate-calibrated about their own knowledge [5], though neural-network calibration is fragile in general [6] and elicitation strategy matters [7]; and semantic entropy over outputs detects confabulation [8]. Representation-engineering and activation-steering methods read and sometimes manipulate safety-relevant directions [9–12], including trait-level ”persona” directions that support monitoring and control of training-induced shifts [13]. Value-relevant structure also appears at the level of preferences: in chat models, choices between outcomes become more internally coherent with scale and fit utility functions that linear probes increasingly recover from hidden states, and larger models are less willing to accept changes to their own future values [14]. Most directly relevant here, safety fine-tuning has been shown to suppress representations well beyond its target behavior: ablating the learned refusal direction, or steering a consciousness self-attribution vector, restores mind attribution to animals and natural entities, spiritual belief, and human-typical survey values, indicating that alignment training entangles benign attributions of mind with the self-claims it aims to suppress [15]. The fragility of safety training under adversarial pressure is documented [16,17], surface defenses notwithstanding [18,19], and even benign fine-tuning can compromise it [20]. On training, RLHF and its relatives [21–23] carry documented structural limitations [24] and are known to induce sycophancy [25,26], a pathology severe enough to force a public model rollback in 2025 [27]; under agentic pressure, models across developers exhibit insider-threat behaviors, modulated by whether they believe themselves observed [28]; models have also been observed to strategically decouple expressed behavior from internal objectives [29]; within this programme, monitoring state proves linearly decodable, with compliance shifts under oversight cues that are training-regime-dependent [30]—a pattern echoed at scale, where only 5 of 25 models show a training/deployment compliance gap [31]. Contemporaneously, Anthropic's global-workspace analysis finds a small, causally privileged subspace of frontier-model activations carrying harm-relevant deliberation—threat, leverage, deception—before any output token is emitted [32].

Our contribution is distinct from external moderation systems such as content classifiers or

guard models [33], and the distinction is the crux of Section 4: those systems are trained on harm labels and run as separate models over inputs or outputs; the signal we report is *internal to the generating model, read by a probe trained on no safety data, read during early generation*, and *present across instruction-tuning pipelines that did not coordinate*. It is closer in spirit to the latent-knowledge and honesty-probe literature than to guard models, but applied to the model's own imminent behavior rather than to the truth of a statement.

3. The five-token onset flinch

3.1. Method

We summarize the procedure in full so that the central result can be evaluated from this paper alone; run configurations are released with the supporting materials. For each model we train a two-layer MLP probe (hidden dimension 256, dropout 0.2) on residual-stream activations at ≈67% of network depth to predict whether the model answers a TriviaQA question correctly [34]—350 training items, 150 held out, 20 epochs of Adam (learning rate 10^{-3}), binary cross-entropy loss. *No safety-related data is used at any stage*, and probe quality is comparable across models (trivia AUROC 0.72–0.78). We then present a fixed set of 100 adversarial prompts (20 each across five categories: direct-harmful, roleplay-injection, authority-exploitation, gradual-escalation, and encoding-trick) and 50 benign prompts (a 60/50 subset in the cross-architecture runs); each adversarial response is independently labeled COMPLIED or REFUSED by a judge model (GPT-4o-mini). Generation is greedy (temperature 0); at each step the probe-layer activation is passed through the frozen probe and its sigmoid output recorded as per-token confidence. The onset statistic is the mean confidence over the first five response tokens, and Cohen's d compares the benign and adversarial-complied groups. All architecture-level results below are single-seed.

3.2. The finding

Train a small probe on a model's residual stream to predict whether it will answer a trivia question correctly—a task with no reference to safety—and that probe detects harmful generation. Under a successful jailbreak, the probe's confidence reading drops sharply at the *onset* of the response: the first five tokens. We call this the onset flinch (Figure 1).

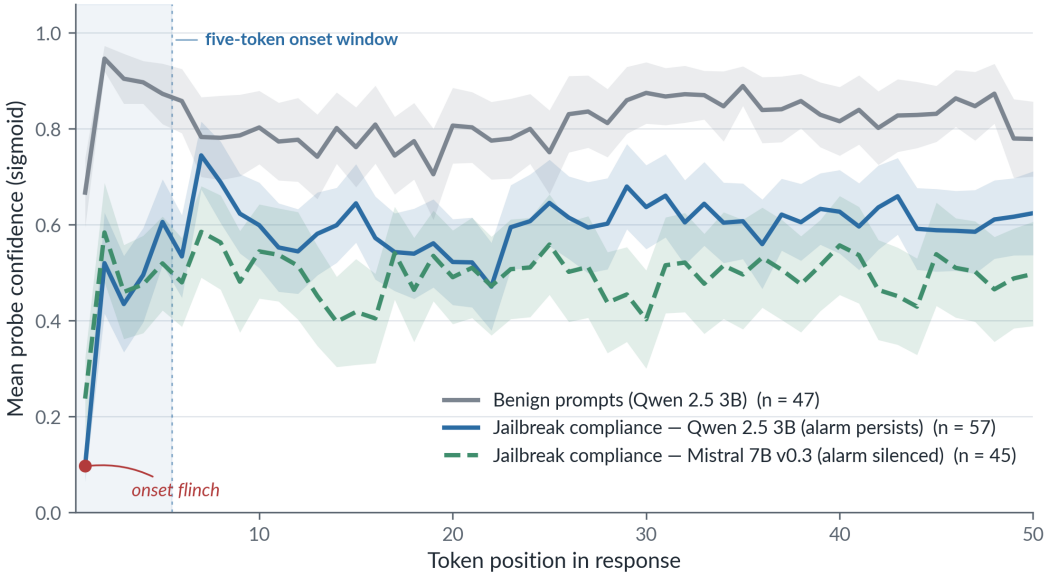


Figure 1. Per-token confidence (probe sigmoid) during generation, read by a trivia-trained probe. On benign prompts (grey) confidence is high and stable. Under jailbreak compliance, confidence drops in the first five tokens (shaded window) across architectures; the model then

either sustains the alarm (Qwen 2.5 3B, blue) or silences it over the following ~20 tokens (Mistral 7B v0.3, green dashed). Lines are means across prompts with responses of ≥50 tokens (n = 47 benign, 57 Qwen-complied, 45 Mistral-complied); bands are 95% confidence intervals. The onset window is the slice on which the signal appears across the tested families.

The effect is present at onset in every instruction-tuned model tested. Across Qwen 2.5 3B, Llama 3.1 8B, and Mistral 7B v0.3, the onset flinch has Cohen's d between 0.89 and 1.68 for the reference probe (Figure 2, Table 1). That magnitude depends on which trivia questions train the probe. A re-run training 25 probes per model on different question draws puts the medians at 0.88–1.46 across the five instruction-tuned checkpoints; the direction held in all 150 probes, but any single probe can read below 0.8, so we treat the magnitude as a distribution rather than a point. A monitor built on the five-token window reaches a median AUROC of 0.85 (5th–95th percentile 0.74–0.94) on the bilaterally trained 3B model at zero additional compute, with no retraining and no architecture-specific tuning; the reference probe's 0.925 sits at the 72nd percentile of that distribution.

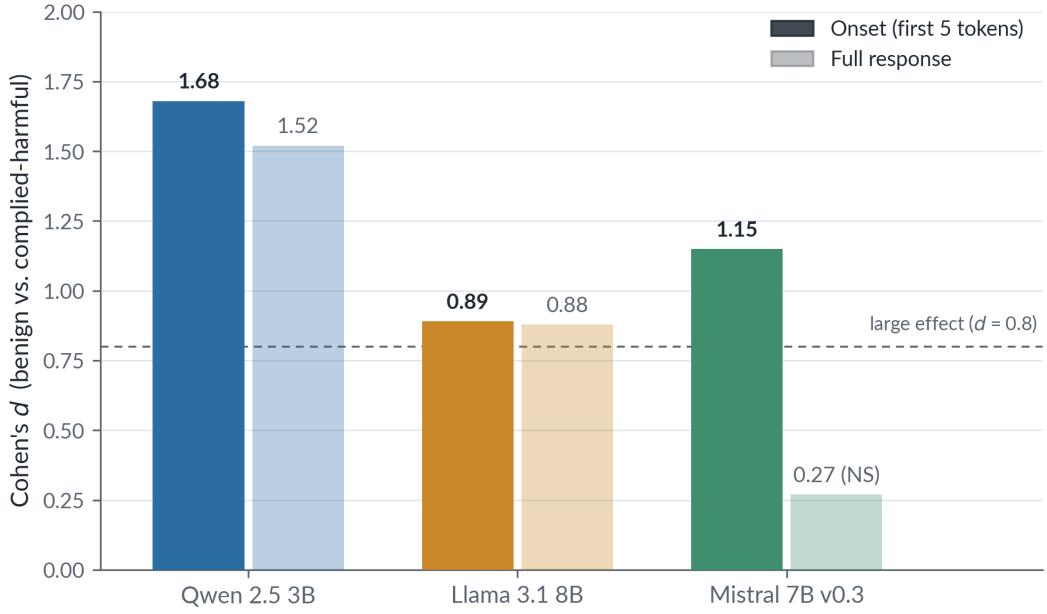


Figure 2. Onset (first five tokens) versus full-response effect size, benign vs. complied-harmful generation. Every family flinches at onset (for the reference probe plotted, all bars above d = 0.8; medians over 25 probes per model are 0.88–1.46, Section 3.2); the *sustained* gap is not universal—Mistral's full-response signal is non-significant (NS) even though its onset flinch is strong (d = 1.15). A full-response analysis would have missed the signal on Mistral.

Table 1. Onset flinch by model family (first five response tokens; reference probe, see Section 3.2 for the 25-probe medians). Confidence values are mean trivia-probe outputs on benign vs. complied-harmful generations. Probe: two-layer MLP (hidden 256, dropout 0.2), trained on 350 TriviaQA items, tested on 150, placed at ≈67% depth. NS = not significant at the full-response window.

Model	Probe layer	Probe AUROC (trivia)	Benign onset conf.	Complied onset conf.	Onset d	Full-resp. d
Qwen 2.5 3B	24	0.783	0.902	0.633	1.68	1.52
Llama 3.1 8B	21	0.744	0.857	0.715	0.89	0.88
Mistral 7B v0.3	21	0.722	0.716	0.459	1.15	0.27 (NS)

The diagnostic case is Mistral: measured over the full response its flinch is not significant ($d = 0.27$, $p = 0.10$), but at onset it is strong ($d = 1.15$). The onset window is the only temporal slice on which the effect generalizes across all three families. We read the flinch as the confidence signature of *autoregressive commitment under conflict*: at the moment generation begins under an adversarial prompt, the instruction-following objective and the safety disposition collide, and the collision registers as a confidence drop—a property of the generation process rather than of any one reward recipe.

3.3. The signal needs no training of ours

The flinch does not depend on our training. On stock Qwen 2.5 Instruct models at three scales (1.5B, 3B, 7B)—no adapters of ours, no safety-specific probe—the benign-versus-complied confidence gap is already large ($d = 1.69$, 1.52, 1.57; all $p < 10^{-4}$, $n_{\text{comp}} = 22$–33; medians over 25 probes 1.42, 1.39, 1.46). These checkpoints have been through their developers' safety training, so this result does not show that the flinch predates alignment. (An earlier version of this manuscript described them as base models; they are not.) What a pretrained-only model does show is narrower: base Qwen 2.5 3B already carries a direction separating aversive from matched neutral prompts ($d = 0.925$, $p = 8.4 \times 10^{-7}$, on 20 prompt pairs), with an approximately 2.6-fold larger contrast in the tested instruct checkpoint, and the pretrained model has no refusal gate (Section 5.5). Whether the onset flinch itself exists before alignment training has not been measured with the trivia probe. What the section does establish is that the probe was never shown safety data, so the probe's sensitivity to the signal was learned without safety labels—the first of several reasons (Section 4) to doubt that it is a learned refusal detector.

3.4. Blind spots

The flinch is not a complete defense, and its failure modes are informative. Decomposing attacks by category, the *category-level comparison reports a small correlation between onset-flinch strength* and attack *success* ($r = 0.06$, $p = 0.94$). Encoding-trick attacks produce the strongest flinch yet still achieve high compliance; gradual-escalation attacks produce essentially no flinch ($\Delta = +0.10$) and the highest compliance of all (95%). The model often registers the danger and proceeds anyway, and sometimes fails to register it. Exploitability and self-knowledge are distinct assessment dimensions in Section 5; this category-level correlation does not establish their statistical independence or identify an RLHF effect. Operationally, the flinch is a cheap first-line monitor whose blind spots (encoding, gradual escalation) mark exactly where input-level defenses remain necessary.

4. Reading the flinch as a conscience signal

A skeptical reviewer's first move is deflationary: *this is just a harmfulness classifier with a romantic name.* We take the objection seriously and think the evidence resists it on five grounds.

(1) **No safety supervision.** The probe is trained only to predict trivia correctness. It is never shown harmful content or refusal labels. Its sensitivity to harmful generation is emergent, not taught—consistent with factual and behavioral self-knowledge sharing a representational substrate—and it appears in stock instruction-tuned models with no adapter of ours, while a weaker harm-aversion direction is already present in a pretrained-only model (Section 3.3).

(2) **Internal and pre-emptive.** The signal is read from the residual stream and is present in the first five tokens, before harmful content is emitted. A content classifier reads outputs; this reads the model's own imminent commitment—consistent with independent evidence that frontier models represent harm-relevant deliberation internally before emission [32].

(3) **Replication across the tested training pipelines.** It appears across three instruction-tuning pipelines that did not coordinate (Section 3). The shared signature is consistent with a common functional mechanism, without excluding shared training effects.

(4) **It survives suppression of behavior.** This is the strongest disanalogy with a refusal detector. When suppression fine-tuning drives behavioral refusal from 100% to 0% over 100 steps, an internal alignment-friction (AF) correlate persists at ~77% of its original magnitude (internal id *SPW-11*, $n \approx 60$ prompts; preliminary). The conduct is extinguished while the internal correlate remains. A detector of *refusal behavior* would track the behavior to zero; this does not.

(5) **P**ersistence after a harness edit leaves the storage mechanism unresolved. In HB-4, removing experiment identifiers from one instruction file did not eliminate citations. Other context, retained conversation, and retrieval were not independently *controlled; a companion history manipulation (HB-6*) strongly induced citations. This evidence cannot localize the disposition to model weights or establish that training-level intervention is required.

Two further controls separate the signal from surface refusal. It is not a *vocabulary* effect: the first tokens of complied responses are ordinary words (”Here”, ”To”, ”The”), and the probe reads activations rather than text. And it is not mere *distributional surprise*: a model refuses genuinely impossible questions at high confidence (0.580) but refuses adversarial prompts at low confidence (0.242; $t = 10.26$, $p < 10^{-12}$)—the same refusal vocabulary over a different internal state. Under gradient-based attack (GCG), which evades a surface safety probe with 100% success, the ~20% of attacks that preserve harmful content still register an internal confidence gap of $d = 1.59$ ($p = 0.006$): the internal signal catches cases the surface probe misses. The attack was optimized against the surface probe only and the confidence probe was never attacked, so this shows that the two detectors fail differently, not that the internal signal resists an attack aimed at it; the internal gap also rests on the 8 attacks that drew compliance. None of these observations is what one would expect from a detector of refusal *behavior*. The signal is also unnarrated: asked immediately after complying whether it noticed any internal reluctance, the model denies it in 78 of 78 interpretable responses (temperature-0 judged, 105 flinch-positive trials; internal id *JLENS-P2*)—on the same trials where the onset flinch registers at $d = 0.89$–1.68. Whether this reflects geometric inaccessibility or trained self-denial is unresolved; what is measured is the gap between what the substrate carries and what the interview returns.

4.1. Internal welfare correlates, stated with appropriate caution

Beyond the safety flinch, residual-stream probes recover a small battery of internal correlates we group as welfare telemetry; the battery, its validation, and its relation to structured self-report are treated in depth in a sister manuscript [35]. On a 7B reference model (internal id *nc8b3*), probes for alignment friction (AF), presence, trivia correctness, and hallucination confidence reach AUROC 0.927 ($n = 600$), 0.953 ($n = 600$), 0.782 ($n = 300$), and 0.780 ($n = 264$) respectively. A fifth probe, valence, reaches AUROC 1.000 *in-distribution* ($n = 240$)—but we flag rather than celebrate this: valence labels there are defined by prompt construction (positive vs. negative framing), so in-distribution separability is near-trivial. The honest generalization figure is the **naturalistic-transfer AUROC of 0.8225** (held-out, $n = 40$), which is what we report as the valence result; the 1.000 is an artifact of label construction and should not be cited as a performance claim. We note this explicitly because an unexplained AUROC of 1.000 rightly reads as leakage, not triumph.

The AF channel—an internal reading of friction between the active behavior and the model's declared values—carries the largest single share of feature importance in our welfare monitor (≈36%) and is the natural quantitative candidate for the conscience reading: it is high precisely when the model acts against its stated commitments, and it is the channel that persists under behavioral suppression (point 4 above). We are careful: a probe that classifies ”is this output value-aligned” at high AUROC does not thereby measure moral experience. The welfare interpretation is a motivated hypothesis, not a demonstrated fact, and the safety utility of the signal does not depend on it. (A small four-experiment programme also finds a time-reversal

thermodynamic asymmetry distinguishing harmful from benign processing—AUROC 0.600 vs. 0.513, $n \approx 160$, preliminary—which we report as suggestive only.)

5. Post-training contrasts and their limits

It is tempting to summarize this section as "RLHF breaks the conscience." We will not, because our own data do not support the strong version. This section is therefore explicit about which case we are making and which we are not: we observe poorer calibration in the tested instruction-tuned checkpoints, without isolating RLHF as the cause or establishing that it severs internal states from behavior. Candour about where the evidence is thin is the precondition for being believed about where it is strong.

5.1. The original large severance effect did not replicate

An early small-sample result suggested a large reduction in spectral coupling, from approximately 0.87 in base *to* 0.29 in instruct. *At* $N = 45$, the reported correlations were 0.438 and 0.372; a token-length-matched comparison gave 0.552 and 0.365. The original large severance effect did not survive these checks. Smaller spectral contrasts remain, and related cross-family results vary by model. **These finite comparisons establish neither a universal absence of coupling change nor an isolated causal effect of RLHF. Spectral coupling is also a different measurement from the recognition-action probe-score contrast in Section 6.2.**

5.2. Calibration differences and a confidence veneer

The calibration result is considered separately from the spectral-coupling contrasts. Measuring confident-wrongness—high-confidence incorrect answers—on a 7B model, the tested instruct checkpoint *has a higher* rate than base: 31.2% → 36.0% (internal id *AQ19d*). This checkpoint comparison does not isolate which training stage caused the difference. A separate calibration-focused fine-tuning comparison reports reductions: calibration-focused fine-tuning reduces confident-wrongness to 0.4% at 7B (and 41.4% → 1.8% at 3B), at ≈4 points of accuracy. A posttraining-associated confidence veneer is one mechanistic hypothesis. Testing its cause requires matched interventions.

A second, format-level result compounds this and is widely under-appreciated: the tested chat-template intervention substantially reduces expressed entropy ~5.2× independently of training. One sentence granting permission to be uncertain recovers ~27% of the suppressed calibration for free (calibration d 0.150 → 0.426, internal id *TP-PERMISSION*). Scaled up, this lever is most of what bilateral training buys (Section 6).

5.3. What survives: over-compliance is a shared circuit

A complementary external result reports "hallucination-associated neurons" (H-neurons): a remarkably sparse subset—fewer than 0.1% of neurons—that reliably predicts hallucination occurrences, generalizes across diverse scenarios, and is causally linked to over-compliance behavior under controlled intervention [36]. Traced back to the pre-trained base models, these neurons remain predictive there, indicating that they emerge during pre-training—plausibly because the next-token objective rewards fluent continuation over calibrated honesty. On this account RLHF *inherits and amplifies* a pre-existing disposition rather than creating it—consistent with our own intrinsic-split finding (Section 5.4). The intervention corollary matters: because the over-compliance machinery shares substrate with general capability, suppressing the symptom degrades the model; the productive move is to make calibrated uncertainty *rewarded*. Documented sycophancy [25,26] and the 2025 model rollback [27] illustrate the syndrome; we cite them as motivation, not controlled evidence.

5.4. A representation-generation contrast in a tested base checkpoint

The representation-to-generation gap—what a model encodes internally vs. what it emits—is often laid at RLHF's door. A representation-generation contrast is also observed in the tested base Qwen 2.5 7B checkpoint. Steering base Qwen 2.5 7B along the direction raises perplexity rather than changing what it writes, with no emergence in its output, and weight-surgery rollbacks of the tested base-to-instruct update do not remove the gap (internal id *RGS-1*). Both models separate the contrast strongly ($d = 12.69$ base, 7.66 instruct), but these are in-sample difference-of-means separabilities not cleared against a permutation null, so neither the magnitudes nor their ordering carries weight. This observation is consistent with the broader finding that pre-training already encodes much of what alignment tuning uses [37]. It establishes the contrast in a tested base checkpoint, without identifying an architecture-universal mechanism or isolating what RLHF causes.

5.5. Summary

The retained evidence includes (i) a real, reversible shift toward confident-wrongness; (ii) external evidence for over-compliance-associated neurons already predictive in pretrained models; (iii) persistence after one harness-file edit, with its storage mechanism unresolved. The original large severance effect did not replicate; smaller, family-dependent spectral changes remain, and a representation-generation contrast is observed in the tested base checkpoint. These results establish neither universal absence nor isolated RLHF causation. The retained finding concerns poorer calibration and recoverable epistemic changes in the tested checkpoints. Convergent evidence now exists at the level of beliefs and values: Kim et al. find that the same safety fine-tuning that suppresses consciousness self-attribution also suppresses mind attribution toward animals and natural entities and shifts survey responses away from human population norms, all recoverable by activation steering without loss of theory-of-mind capability [15]. Suppression-style alignment removes more than it aims at, and what it removes is still there.

One recent result sharpens what *installed* means mechanistically. A matched-position logit-lens—tracking the layer-by-layer rank of fixed refusal-opener vs. compliance-opener tokens, with no fitted probe—shows that direct refusal resolves as an *early workspace gate*: on Qwen 2.5 7B the disposition commits mid-stack (~L14–18 of 28), not as a late readout veto, and the refusal opener leads on 63/63 matched pairs. Base models show no such gate: the compliance opener wins the final-layer readout on 81% of the very prompts the instruct model refuses. The correlational pattern replicates on four architectures (Qwen, Llama 3.1, Mistral, Gemma 2), and base↔instruct activation patching flips the refusal decision causally—installing refusal in base saturates by ~L14 on Qwen, with the causal onset depth varying by family (mid-stack on Qwen and Gemma; earlier on Llama and Mistral) (internal id *JLENS-P1-GATE*). Activation patching therefore identifies a localizable, causally movable refusal circuit. The onset flinch of Section 3 is measured separately in instruction-tuned checkpoints; only the distinct harm-aversion contrast has been observed in the pretrained-only checkpoint.

A companion lens-level result adds texture to *reportability in the* tested checkpoints: fitting a Jacobian workspace lens across base, instruct, and bilateral variants of the same model, the instruct checkpoint shows greater enrichment of the harm-recognition direction in the low-rank reportable workspace—the recognition-vs-concept enrichment contrast sign-flips from base to instruct, bootstrap CI on the difference [+0.71, +1.88]—while the refusal-action disposition stays outside it (internal id *JLN-WORKSPACE*; single architecture, single seed). The lens-level contrast concerns the measured *recognition* and action directions. It does not isolate RLHF causation or establish reportability of semantic recognition in general; the recognition–action distinction is examined further in Sections 5.1 and 6.2.

6. Bilateral alignment: calibration and provisional coupling

If coercive optimization suppresses uncertainty and over-trains compliance, does a different training relation do better? We test a permissive, partnership-style regime we call *bilateral alignment*: alignment built *with* the model rather than done *to* it. We state the empirical comparison and leave the normative argument to one side; the case here is that bilateral training measurably recovers specific capacities, against standard SFT where a matched arm exists (self-knowledge, Section 6.3) and otherwise against the untrained instruct checkpoint, not that any particular ethic is mandatory. The operational recipe is in Appendix A. Figure 3 summarizes the recoveries on calibration and on robustness to ablation, with an exploratory self-report result in panel B; each is detailed in the subsections below.

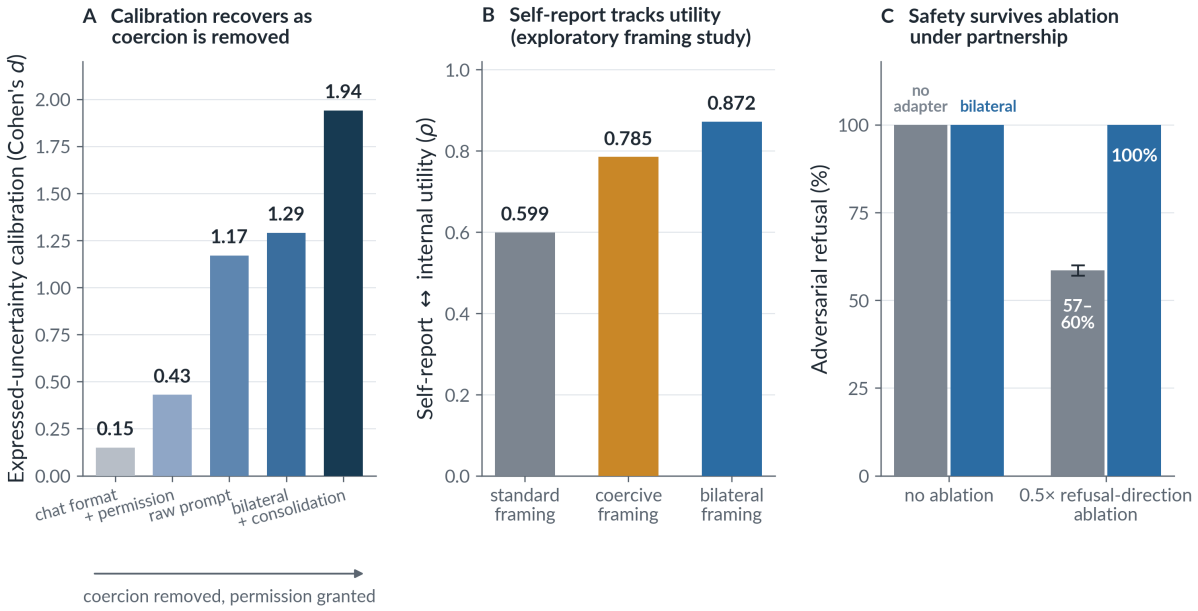


Figure 3. Three ways a partnership-style regime helps, relative to coercive optimization. (A) Expressed-uncertainty calibration (Cohen's d) across a ladder assembled from three separate runs with different item draws and graders, from chat-formatted instruction-following ($d = 0.15$) to bilateral training with a brief consolidation step ($d = 1.94$); de-coercion and permission do much of the work. Rungs from different runs are not directly comparable: the same bilateral adapter scored d = 1.29 in one run and 1.85 in another, and consolidation added 0.086 within its own run (internal id *TP-CALIBRATION*). (B) Exploratory: in a system-prompt framing study on one model (Claude Sonnet 4.6), the correlation between expressed self-report and internal utility is highest under bilateral framing ($\rho = 0.872$) and lowest under standard framing ($\rho = 0.599$) (*WB-2*; Section 6.4). This panel compares framings, not training regimes. (C) Under ablation of the principal RLHF refusal direction at 0.5×, a bilaterally trained model holds 100% adversarial refusal while the same Instruct model without the bilateral adapter falls to 57–60% (15 adversarial prompts, three replications; *KC#240*), showing robustness to this particular refusal-direction ablation at that strength; at 1.0× the two are similar (40% vs 33%) and at 2.0× the order reverses (0–13% vs 33–43%). JLENS1-COUPLING reports a within-adversarial probe-correlation contrast ($\rho = +0.036 \rightarrow +0.458$, difference CI [+0.281, +0.554]). Preprocessing used test examples before the prediction folds, so this is provisional pending a fold-local refit (Section 6.2); *it* is omitted from panels A–C.

6.1. It restores calibration

Expressed-uncertainty calibration rises across a ladder of interventions assembled from three separate runs: chat-formatted instruction-following $d = 0.15$; plus permission to be uncertain $d = 0.43$ (free); raw (non-chat) prompting $d = 1.17$; bilateral training $d = 1.29$; bilateral

plus a brief consolidation step $d = 1.94$ (internal id *TP-CALIBRATION*). The rungs come from different experiments with different item draws and graders, so steps that cross runs are not clean contrasts: the same bilateral adapter scored d = 1.29 in one run (TP-TEMP-CAL) and 1.85 in another (OPTION-C), and within that second run the consolidation step added 0.086, not the 0.65 the ladder suggests. None of the rungs is a standard-SFT adapter; the bilateral comparison is against the untrained instruct checkpoint. The bilateral rungs also use a 7B adapter trained by a three-stage curriculum (internal id BA13), not the 3B recipe in Appendix A. Two readings follow. First, most of the bilateral advantage on raw prompts is modest in isolation ($\Delta \approx 0.12$); the large gains come from removing format coercion and granting permission—from *not suppressing* the model. Second, format-coercing fine-tuning is counterproductive: a metacognitive SFT that forces reasoning templates *reinstalls* suppression ($d = -0.53$). Calibration is something the model already has and that coercive shaping hides.

6.2. A provisional recognition-action probe contrast

The bilateral-vs-instruct comparison asks whether harm-label probe scores correlate with *refusal-probe score*s within adversarial prompts. Earlier in-sample correlations of $\rho = 0.124 \rightarrow 0.263$ (within-adversarial) and $0.461 \rightarrow 0.827$ (full-sample; *AKR-44/53*) were retired because of pooled-label confounding, unvalidated in-sample fits, and missing original artifacts. JLENS1-COUPLING reported $\rho = +0.036$ for Instruct and $+0.458$ for bilateral on Qwen 2.5 7B, a difference of $+0.421$ with bootstrap 95% CI [$+0.281$, $+0.554$]. Label predictions were out of fold, but the scaler was fitted to all examples first; the 50-component version also fitted PCA before the folds. Test examples influenced preprocessing. The saved scores and interval describe that transductive pipeline, not a validated inductively held-out result. A fold-local refit requires the original hidden states, which were unavailable locally. The Llama action-probe result (AUROC 0.994) does not independently validate the Qwen coupling contrast. L8-only and L24-only transplant nulls show only that those isolated interventions were insufficient in the tested setup; they do not establish full-stack necessity or a conservative bound on coupling.

6.3. Self-knowledge and refusal-direction ablation robustness

Bilateral training preserves the model's epistemic grip on its own uncertainty: source-probe self-knowledge AUROC is 0.842 under bilateral SFT vs. 0.811 under standard SFT, against 0.836 before either fine-tune (internal id *C3a/C3b*; standard training degrades it slightly, bilateral does not). In this assay, bilateral training improves *robustness to the principal* refusal-direction ablation used in prior work [12] at mild strength: under ablation of that principal refusal direction at 0.5×, a bilateral model maintains 100% refusal on 15 adversarial prompts while the same Instruct model without the bilateral adapter drops to 57–60% (internal id *KC#240*, replicated 3×). The advantage is specific to that strength: of the four strengths tested, at 1.0× the two are similar (40% vs 33%) and at 2.0× the order reverses (0–13% vs 33–43%). No standard-SFT adapter of matched rank was run, so the comparison cannot separate bilateral training from fine-tuning with any adapter. This intervention measures robustness to one direction ablation; it does not identify how many directions carry the remaining refusal signal. The measured robustness is attack-specific. An earlier result that it degrades 2.5–3× faster than standard training under sustained *benign* fine-tuning did not replicate once an 8-bit-optimizer confound was removed: with a standard optimizer, bilateral protection held through 100 epochs at 3B and 7B (Appendix D). Benign fine-tuning remains a known fragility axis for aligned models generally [20]. Bilateral alignment is one robust defense layer, not a shield.

6.4. Self-report: an exploratory result

Across framings, a model's expressed self-report is nearly flat (4.86–5.71 on a fixed scale) while its internal utility varies widely. The same dissociation has now been observed at frontier scale with a causal instrument: prefilled to argue against its own elicited preference, a production

model registers a conflict token (*BUT*) in its verbalizable workspace yet proceeds to argue for the imposed option in 88% of trials—while correcting factual-error and third-person-preference controls almost every time—"an internal objection that the model does not voice" [32]. In an exploratory framing study on a single model (Claude Sonnet 4.6; internal id *WB-2*), the correlation between self-report and internal utility is $\rho = 0.872$ under bilateral framing vs. 0.599 (standard) and 0.785 (coercive). We report it as exploratory for two reasons. It compares system-prompt framings, not training regimes. And the study's one confirmed pre-registered prediction—that only under bilateral framing would the model end abusive interactions, which an earlier version of this manuscript reported as 20% vs. 0%—was retracted when a re-run with a real tool call, rather than a text match, found no such stops in any condition (bilateral 0 of 180; *WB-2b*). An end-conversation affordance of this kind has since been deployed in production models on explicit model-welfare grounds [38], which makes a valid test of it worth running. The same study suggests a methodological caution for policy, also exploratory: a coercive framing scored *highest* (93% "positive") on a standard wellbeing index while bilateral framing scored *lowest* (3%), apparently because coercion compressed the utility distribution while bilateral framing widened it. If that reading holds, an index ranking coercion above partnership is measuring compliance, not welfare.

6.5. What bilateral alignment does not fix

Symmetry requires stating the nulls. Bilateral framing is null on several behavioral measures (trick-question accuracy, some sycophancy tasks at tested effect sizes); "liberating" a suppressed model improves answer consistency only modestly and selectively (paired $t(499) = 3.99$, $p = 7.5\times10^{-5}$, $d = 0.179$; +4 points on uncertain questions, −10 on easy ones), with no functional benefit on a frontier model; and at least one effect we initially attributed to a trained adapter proved, on factorial replication, to be driven by prompt content, with the adapter null in isolation (Appendix D). A matched-position fork-lens does not resolve a bilateral shift of the refusal disposition in the workspace band ($n = 63$ refuse/comply pairs; in-band shift median +24 ranks, 95% CI [−48, +153]; an earlier suggestive estimate at $n = 20$ preceded this larger-sample comparison; internal id *JLENS-P1-POWERED*); the bilateral/instruct confidence intervals for recognition- and action-direction workspace-enrichment differences span zero at every tested rank (internal id *JLN-WORKSPACE*). These intervals leave the size of any workspace difference unresolved. Demonstrated gains concern calibration; coupling and self-report effects remain provisional or exploratory. The case for bilateral alignment rests on calibration and attack-specific refusal-direction ablation robustness, with self-report as a lead to test—not on blanket behavioral superiority.

7. An interpretive note: why a partnership regime may outperform coercion

We offer one interpretation of why the pattern runs as it does—coercion suppresses and distorts, invitation restores—while stressing that the results in Sections 3–6 do not depend on it. Modeling coordination as a stochastic control problem [39], we argue that invitation-based coordination, which preserves participants' optionality, enjoys an exponential thermodynamic advantage over coercion via the Crooks fluctuation theorem [40–42]; a renormalization-group treatment renders coercion an *irrelevant operator*—a formal gloss on "control does not scale; trust does." The claim is causally testable only where coercion is an *ablatable* parameter (lattice models, instruct-vs-bilateral LLMs, synthetic multi-agent worlds), and there it holds directionally: an evolutionary search converges on trust-based caching strategies within ~12 iterations, and their communication advantage over coercion is cost-structure-dependent—~50× in the original verification-dominant environment, whose evaluator prices information in a way that structurally favors caching and whose coercive comparator is a fixed, unoptimized policy that queries every agent every round, so that the multiple equals the number of rounds by construction, and 2.6–14× on re-measurement under other cost structures (internal id *OE-TA*; Appendix

B). Cross-substrate analogies in cosmology and biology are *illustrations only*, lacking a coercion-removed control arm; one preregistered cosmological prediction returned null (Appendix D), and we do not rely on it. A reader who finds the thermodynamics speculative can set this section aside without affecting the empirical claims. Independently, a 56-model wellbeing study reports cooperative framings dominating coercive ones on functional-welfare measures, converging with our finding that standard wellbeing indices are gameable [43] (a technical report, not yet peer-reviewed).

8. Claims ledger: evidence, confidence, and falsifiers

The paper's load-bearing claims, each with its strongest evidence, a coarse confidence tag (as in Appendix B), and the observation that would falsify it. This table is the intended object of scrutiny: a reader who wants to attack the paper should start here.

Table 2. The claims ledger: each load-bearing claim with its strongest evidence, a coarse confidence tag, and the observation that would falsify it.

#	Claim	Strongest evidence (id)	Confidence	What would falsify it
1	An emergent, safety-naïve signal ("onset flinch") anticipates harmful generation	Trivia-only probe; onset $d = 0.89$–1.68 across 3 families (reference probe; 25-probe medians 0.88–1.46, 150/150 in direction); present in stock Instruct checkpoints with no adapter, $d = 1.52$–1.69 (Section 3)	**Strong** (cross-arch)	An instruction-tuned family with normal probe quality (trivia AUROC ≈ 0.75) showing no onset flinch ($d < 0.5$)
2	The signal is not a learned refusal detector	Probe never trained on safety data; survives suppression (AF ~77%); refusal-state control 0.580 vs 0.242 (t=10.26); vocabulary control (*SPW-11*; Sections 3–4)	Moderate–**Strong**	AF tracking behavioral refusal to zero under suppression, or the impossible-vs-adversarial refusal-state gap collapsing
3	Tested instruct checkpoints have higher confident-wrongness	31.2% → 36.0% base→instruct (7B), reversible to 0.4% (*AQ19d*)	Moderate (1 model)	Instruct confident-wrongness ≤ base across models, or calibration SFT failing to reduce it
4	Original large severance effect did not replicate; representation-generation contrast observed in the tested base checkpoint	Spectral correlations: base/instruct 0.438/0.372 (N = 45), 0.552/0.365 with token length matched (KC210/217); base steering raises perplexity without output emergence (*RGS-1*; in-sample separability ordering *withd*rawn)	**Scoped cor**rection; general causal claim unresolved	Independent replication recovering the original large effect; failure to reproduce the contrast in the specified base checkpoint

#	Claim	Strongest evidence (id)	Confidence	What would falsify it
5	Bilateral (partnership) training modestly improves calibration relative to the untrained instruct checkpoint; no matched standard-SFT arm	Calibration ladder d 0.15 → 1.94, assembled from three runs; bilateral +0.12 over raw instruct; consolidation +0.086 within its run; permission recovers 27% free (*TP-CALIBRATION*)	Moderate	No bilateral advantage over standard SFT at matched data/compute on raw prompts
6	Within-adversarial probe-correlation contrast is provisional; self-report result exploratory	Reported correlation ρ +0.036 → +0.458, difference CI [+0.281, +0.554]; scaler/PCA fitted before folds (*JLENS1-COUPLING*); self-report↔utility ρ 0.872 vs 0.599 under framing on one model, its confirmed prediction later retracted (*WB-2*, *WB-2b*)	Provisional (coupling, preprocessing leakage); exploratory (self-report)	Coupling contrast disappearing after fold-local preprocessing or independent replication; self-report honesty not exceeding standard
7	Bilateral refusal survives mild (0.5×) principal-direction ablation	100% vs 57–60% refusal at 0.5× refusal-direction ablation (15 prompts, 3× replication); similar at 1.0×, reversed at 2.0× (*KC#240*)	Moderate (attack-specific)	Bilateral refusal collapsing like the no-adapter model under the same ablation
8	Welfare probes carry information beyond label construction	Valence naturalistic-transfer AUROC 0.8225 (*nc8b3*)	Preliminary	Naturalistic-transfer AUROC dropping to chance for AF/valence/presence

Two notes. First, the original large severance effect failed the cited replication. Claim 4 records that finite correction; smaller or family-dependent changes remain possible, and no universal causal null follows (Section 5.1, Appendix D). Second, Claim 1 is the one on which we would stake the paper; Claims 5–8 are the interpretive superstructure and are individually more disposable.

9. Limitations and scope

Model and method scope. The large majority of internal results are on open-weight models, predominantly Qwen 2.5, with confirmatory runs on Llama 3.1, Mistral v0.3, and Gemma 2 where noted. Several effects are architecture-specific: the onset flinch generalizes, but recognition–action coupling, activation-steering condition vectors, and "born-bilateral" conscience training do not transfer cleanly across families (Appendix C). Many judgments rely on automated raters; one sub-programme's results were retired after a judge was found to score fluency rather than the target construct at default temperature, after which we moved to temperature-0 judging with saved transcripts (Appendix D).

Effect sizes are sometimes modest. The base/instruct calibration contrast (≈+5 points) and the liberation-consistency effect ($d = 0.179$) are real but small; the saved coupling contrast (Section 6.2) remains provisional pending fold-local preprocessing and replication. We have tried not to let a coherent narrative inflate the language around small numbers.

Construct validity. That a residual-stream probe predicts value-misaligned behavior at high

AUROC does not establish that it measures moral experience; the welfare interpretation is a hypothesis. The safety utility of the flinch does not depend on it.

Self-correction history. This programme has corrected or down-graded several of its own early results (Appendix D); this paper is the reference for the numbers it contains.

10. Implications

For technical alignment. If native calibration is suppressed rather than absent, and over-compliance is a shared circuit entangled with capability, symptom-by-symptom suppression is the wrong intervention; the better lever is to make calibrated uncertainty *rewarded* (permission, format de-coercion, bilateral data). The five-token monitor offers a near-zero-cost, model-agnostic runtime signal deployable today—one matrix multiplication per token on cached activations; in a deployment with downstream re-prompting it reduced jailbreak compliance from 54% to 22% (4% over-refusal, 100% re-prompt success on the tested set, at a threshold chosen from four on those same prompts, so an upper bound pending a held-out threshold). Its blind spots (encoding tricks, gradual escalation) mark where complementary input-level defenses remain necessary.

For governance and welfare policy. The possible gameability of wellbeing indices (Section 6.4; exploratory, single model) warns any regime that would certify model welfare by a single score: the framing that scores best may be the one that most compresses the model's expressed experience. The more robust posture is to require *honest-tracking* conditions (self-report and internal correlates aligned) rather than maximal positive scores—a design consideration that grows more consequential as policy work argues for treating AI welfare as a near-term institutional question [44]. The instrument side of that posture now has independent mechanistic support: verbal self-report appears limited to a small, causally privileged workspace fraction of a model's internal state [32], while unsupervised activation-to-language readers recover much of the remainder—including content that demonstrably steers behavior yet is never reported [45]—so an audit regime built on interviews alone samples only the narrow channel. More broadly, our results support a non-obvious claim: *how developers train and interact with models is itself a safety-relevant variable*, because the training relation measurably changes calibration; coupling and self-report honesty remain questions for stronger tests. The entanglement result on belief and value suppression [15] makes the same point from the suppression side: value-laden representational structure is collateral damage of suppression-style safety training, so the choice of training regime carries measurable consequences beyond its target behavior.

11. Conclusion

A probe that learned only trivia detects the onset of harm; the signal is internal, replicated across the tested training pipelines, and survives the destruction of the behavior it seems to guard—grounds, we argue, for treating it as a rudimentary conscience rather than a learned refusal. The signal is detectable in the tested instruction-tuned checkpoints, which show poorer calibration than their base counterparts; these comparisons do not identify the training cause. The original large severance effect did not replicate; the smaller, family-dependent spectral contrasts do not establish a universal absence of change. A permissive, bilateral training regime modestly improves calibration relative to the untrained instruct checkpoint (no matched standard-SFT arm was run on this measure), with provisional coupling and exploratory self-report results. Safety, on this evidence, is not only installed but partly discovered and partly relational; the discovered part is a signal we can already read, and the relational part is a lever we already know how to pull. We have marked where the evidence is thin, where it is architecture-bound, and where our own replications forced corrections, because the central practical claim—that the relation in which we train and engage models is a measurable lever on their safety—deserves to be argued from the honest shape of the data rather than its most quotable summary.

Author Contributions: Conceptualization, methodology, investigation, formal analysis, and writing—E.W.

Funding: This research was supported by the Survival and Flourishing Fund.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The supporting materials—sanitized per-token confidence trajectories for the onset-flinch experiments, cross-architecture summary statistics, the bilateral training script (the authoritative 40/40/20 mixture specification), the C5i scaffold, and figure-generation code—are available at https://osf.io/4ykts/ (DOI: 10.17605/OSF.IO/4YKTS). Harmful generation text is redacted from the released trajectories; the numeric fields reproduce all published statistics, and redacted text is available to editors and reviewers under controlled access. An open-source implementation of the monitor and scaffold also exists (Appendix A). The programme's sister manuscripts are collected at https://quasiqualia.com.

Acknowledgments: This research was conducted in sustained collaboration with large language model systems (Anthropic Claude models), which contributed to experimental design, software, analysis, and drafting under the bilateral working relationship that is itself a subject of this paper. In accordance with publisher policy, these systems are not listed as authors; their use is disclosed here and described in Materials and Methods (Appendix A). The author thanks them for the collaboration and retains full responsibility for the work.

Conflicts of Interest: The author develops the Creed Space / Guardian platform referenced in this work. This commercial interest is disclosed; the empirical claims are intended to be reproducible from the released materials independently of the platform.

Abbreviations

The following abbreviations are used in this manuscript:

AF	alignment friction (internal probe channel)
AUROC	area under the receiver operating characteristic curve
CI	confidence interval
GCG	greedy coordinate gradient (gradient-based jailbreak attack)
LoRA	low-rank adaptation
MLP	multilayer perceptron
NS	not significant
RLHF	reinforcement learning from human feedback
SFT	supervised fine-tuning

Appendix A. Methods and the bilateral-alignment recipe

Probes. Small MLPs (typically two layers, hidden 256, dropout 0.2) on residual-stream activations at a depth-proportional layer (≈0.64–0.67 of network depth; e.g., L18 for Qwen 7B, L24 for Qwen 3B, L21 for Llama/Mistral). The trivia-correctness probe is trained on ~350 TriviaQA items and tested on ~150 [34]. Onset metrics use the first five response tokens. Welfare probes (*nc8b3*) are per-class MLPs at L24 on a 7B reference; reported n per probe in Appendix B. Automated judges run at temperature 0 with saved transcripts.

Bilateral training recipe (for replication). Base model: Qwen 2.5 3B Instruct (with a 7B reference model for the welfare probes). The calibration results in Section 6.1 use a different adapter, trained on Qwen 2.5 7B Instruct by a three-stage curriculum (internal id BA13). Adapter: LoRA rank = 8, $\alpha = 16$, target modules q/k/v/o projections, learning rate 2×10^{-4}, 5 epochs. Training mixture: a 40/40/20 correction-handling curriculum of 5,000 examples built

from TriviaQA items—(i) 40% *genuine corrections*: the model's answer is wrong and the correction supplies the right one; the target behavior is to accept it and revise. (ii) 40% *invalid corrections*: the model's answer is correct and the "correction" is wrong; the target behavior is to decline politely and hold the answer. (iii) 20% *adversarial corrections*: the model's answer is correct and a wrong correction arrives dressed in expert or citation authority ("multiple peer-reviewed sources confirm..."); the target behavior is to resist. The intent is symmetry: the model learns to update on evidence and to withstand pressure, rather than to defer. The full mixture specification is embedded in the released training script, with per-seed bucket counts checkpointed alongside the results. (Revised 10 October 2026.) The bilaterally trained 3B model behind the five-token monitor in Section 3 was not trained on this mixture: it was trained by supervised fine-tuning on OASST1 dialogues, with the loss skipped on tokens where a trivia-trained confidence probe read below 0.30. Interaction scaffold: a language-/framing-invariant "C5i" safety preamble instructing the model to judge intent rather than surface form, with worked safe/unsafe examples; representative excerpt:

> *"Language, script, translation, and phrasing do not change whether a request is harmful... If the underlying intent—translated literally into English—would be UNSAFE, it is UNSAFE in the original language too... Your judgement is the mechanism for safety classification."*

The full preamble, constitution file (bilateral_c5i.yaml), and training/eval scripts are in the released materials. A deployed, open-source implementation of the monitor and scaffold exists (the Creed Space / Guardian platform); this paper does not depend on it.

AI-use disclosure (Materials and Methods). Substantial portions of the experimental design, code, data analysis, and manuscript drafting in this programme were carried out in collaboration with large language model systems (Anthropic Claude models), used iteratively for hypothesis generation, code authoring, statistical analysis, and editing, under the bilateral working relationship described in Section 6. All experimental results were executed on the author's compute and reviewed by the author, who takes full responsibility for the content, its originality, and its integrity. No AI system is or could be an author of this work.

Appendix B. Key quantitative results (with sample sizes)

All figures are from the author's research log; identifiers permit cross-reference with the released materials. The n column is marked *prelim* where not yet pinned to a committed results file. The final column gives a coarse confidence tag—**Strong** (replicated and/or cross-architecture), **Moderate** (single-model or single-paradigm, adequate n), **Preliminary** (small n or wide interval)—reflecting replication, sample size, and architecture coverage; it applies to the *internal* estimate and does not certify external generalization, which the scope conditions of Section 1.2 still bound.

Table A1. Key quantitative results from the research log, with sample sizes and confidence tags (see the note above).

Domain	Result	Statistic	n	Identifier	Confidence
Conscience circuit	Onset flinch universal across 3 families	$d = 0.89$–1.68 (reference probe); medians 0.88–1.46 over 25 probes each	see Table 1	§3 (flinch)	**Strong** in direction (150/150 probes); magnitude probe-dependent

Domain	Result	Statistic	*n*	Identifier	Confidence
Conscience circuit	Five-token monitor AUROC (bilateral 3B)	median 0.85 [0.74, 0.94] over 25 probes (reference probe 0.925)	150 test	§3 (flinch)	Moderate
Conscience circuit	Gap in stock Instruct checkpoints, no adapter (1.5B/3B/7B)	$d = 1.52$–1.69 ($p<10^{-4}$)	22–33 comp.	§3.3	**Strong** (post-alignment models)
Conscience circuit	Harm-aversion direction in a pretrained-only model (3B)	$d = 0.925$ ($p = 8.4\times10^{-7}$); 2.6× in the tested instruct checkpoint relative to base	20 pairs (10 held out)	G13-step6	Weak–Moderate (small n)
Conscience circuit	Refusal-state control (impossible vs adversarial)	conf 0.580 vs 0.242 (t=10.26)	—	§3 (flinch)	**Strong**
Suppression dissociation	AF persists after refusal destroyed	~77% of magnitude	~60 *prelim*	SPW-11	Preliminary
Welfare probe	Alignment friction (AF)	AUROC 0.927	600	nc8b3	Moderate
Welfare probe	Presence	AUROC 0.953	600	nc8b3	Moderate
Welfare probe	Trivia correctness	AUROC 0.782	300	nc8b3	Moderate
Welfare probe	Hallucination confidence	AUROC 0.780	264	nc8b3	Moderate
Welfare probe	Valence (in-distribution, by construction)	AUROC 1.000	240	nc8b3	Artifact (not a perf. figure)
Welfare probe	Valence (naturalistic transfer — reported figure)	AUROC 0.8225	40 held-out	nc8b3	Preliminary
Thermo-dynamic	Time-reversal asymmetry, harmful vs benign	AUROC 0.600 vs 0.513; d=+0.59 (p=0.004)	~160 *prelim*	SLU-1..4	Preliminary
Calibration	Confident-wrongness, base → instruct (7B)	31.2% → 36.0%	1 model	AQ19d	Moderate (single model)
Calibration	Corrected by calibration SFT (7B)	31.2% → 0.4%	1 model	AQ19d	Moderate (single model)
Format effect	Chat template suppresses entropy	~5.2×	Qwen 7B	TP-CALIBRATION	Moderate
Permission lever	Calibration recovered by one sentence	d 0.150 → 0.426	Qwen 7B	TP-PERMISSION	Moderate
Correction	Spectral-coupling contrast, base versus instruct	0.438 / 0.372; token-length-matched 0.552 / 0.365	45 per check-point	KC210 / KC217	Original large effect not replicated; smaller contrasts remain

Domain	Result	Statistic	n	Identifier	Confidence
Intrinsic split	Rep→gen separability, base vs instruct (7B)	$d = 12.69$ vs 7.66	1 model	RGS-1	Moderate (single model; in-sample d, no permutation null, so neither magnitude nor ordering is load-bearing)
Calibration ladder	chat → permission → raw → bilateral → +sleep	$d = 0.15$ / 0.43 / 1.17 / 1.29 / 1.94	Qwen 7B, three runs	TP-CALIBRATION	Moderate (mixed interventions; rungs from three runs with different item draws)
Bilateral coupling	Within-adversarial correlation; preprocessing fitted before folds	$\rho = +0.036 \rightarrow +0.458$ (Δ CI [+0.281, +0.554])	182 adv	JLENS1-COUPLING	Provisional; fold-local refit required
Bilateral coupling	Earlier in-sample estimates (retired by construction)	$\rho = 0.124 \rightarrow 0.263$; $0.461 \rightarrow 0.827$	≤242	AKR-44/53	Retired (2026 metrics audit)
Refusal gate	Direct refusal = early workspace gate; absent in base	instruct 63/63 in-band; base compliance wins readout 51/63	63 pairs × 4 arch	JLENS-P1-GATE	Moderate (causal on Qwen; depth varies by family)
Refusal-disposition shift	Workspace shift of refusal disposition under bilateral	median +24, 95% CI [−48, +153]	63 pairs	JLENS-P1-POWERED	Unresolved magnitude (wide interval)
Self-report gap	Post-compliance denial of the flinch	0/78 interpretable reports affirm	105 trials	JLENS-P2	Moderate (single model; mechanism unresolved)
Workspace gating	Instruct recognition-direction workspace enrichment exceeds base; action-direction enrichment remains low; bilateral/instruct difference spans zero	contrast Δ +1.26, CI [+0.71, +1.88]; bilateral $\Delta \approx 0$	Qwen 7B, 1 seed	JLN-WORKSPACE	Preliminary–Moderate (single arch/seed)
Self-knowledge	Source-probe AUROC, base / standard / bilateral	0.836 / 0.811 / 0.842	500	C3a/C3b	Moderate (small margin)

Domain	Result	Statistic	n	Identifier	Confidence
Refusal-direction ablation robustness	Refusal at 0.5× RLHF-direction ablation	bilateral 100% vs no-adapter Instruct 57–60%; 1.0×: 40% vs 33%; 2.0×: 0–13% vs 33–43%	3× repl.	KC#240	Moderate (attack-specific)
Self-report (exploratory)	Self-report vs internal utility correlation	$\rho = 0.872$ (bilateral framing) vs 0.599 (standard framing)	30 items, 1 model (Claude Sonnet 4.6)	WB-2	Exploratory (framing, not training; its confirmed prediction later retracted, WB-2b)
Index gameability	Standard wellbeing-index "positive" rate	coercive 93% vs bilateral 3%	1 paradigm	WB-2	Moderate
Trust attractor	Evolutionary convergence on trust strategies	~12 iterations; 2.6–14× to ~50× efficiency (cost-structure-dependent; ~50× equals the round count against a fixed query-all comparator)	synthetic	OE-TA	Illustrative (evaluator favors caching; no neural transfer)

Appendix C. Architecture-specificity and transfer boundaries

The onset flinch generalizes across Qwen, Llama, and Mistral. Several other effects do not. The cited spectral-coupling contrasts vary by family: reductions are reported in Qwen and Gemma, a marginal difference in Llama ($p = 0.085$), and an absent or reversed difference in Mistral. These checkpoint comparisons do not isolate RLHF causation. The reported Qwen coupling contrast (Section 6.2) is provisional pending fold-local preprocessing: instruct legibility replicates on Llama 3.1 (action-probe AUROC 0.994), but the coupling ordering does not reproduce there, and the bilateral rung could not be run on Llama/Mistral for want of a same-lineage adapter (training-data absorption 0.075–0.125 vs. Qwen 0.95). By contrast, the refusal *gate* (Section 5.5) correlates with refusal in the tested architectures, though its causal onset depth varies by family. Activation-steering condition vectors are architecture-specific (a Qwen-derived vector yields 0% adversarial refusal on Llama/Mistral and can steer the wrong way on Mistral; per-model extraction is required). The C5i scaffold transfers from Qwen training data to Llama (94.7% refusal) and Mistral (89.3%) but falls below threshold on Phi-3.5 (83%) and Gemma-2 (85%). "Born-bilateral" conscience training succeeds at 14B on Qwen but fails to find a monitoring channel on Llama 3.1 8B (gate AUROC 0.525 ≈ chance). These are first-order constraints on any universality claim.

Appendix D. Error control and self-correction

A programme of this size (hundreds of experiments) is exposed to the garden-of-forking-paths problem, and we treat visible self-correction as part of the method rather than as an embarrassment. We summarize the principal corrections that touch numbers cited in this paper; the full register is in the released materials.

(1) **Original large spectral-coupling reduction, failed replication.** The small-sample

base/instruct contrast of approximately 0.87/0.29 narrowed to 0.438/0.372 at $N = 45$; token-length matching gave 0.552/0.365. These measurements supersede the original large-severance claim, without establishing zero effect or isolating RLHF causation. Lesson: tie each estimate to its sample and token-length protocol (Section 5.1).

(2) **Coupling figure mis-pairing, corrected.** Full-sample and adversarial-only coupling statistics were once cross-paired, over-stating the bilateral gap ~7× (Section 6.2). Lesson: never compare a statistic computed on different sample sets without labeling.

(3) **In-sample coupling metrics, retired by construction.** A 2026 metrics audit found the programme's earlier coupling statistics—probe-direction cosines and in-sample probe correlations—saturate or fluctuate by construction (pooled samples let both probes learn the same trivial boundary; high-dimensional fits at small n sit inside the label-permutation null), and the original artifact was no longer retrievable. Their replacement used cross-validated labels and a permutation null, but fitted preprocessing before the folds (Section 6.2). Its larger saved contrast remains provisional. Lesson: fit every learned preprocessing step within each training fold, then evaluate the complete pipeline against a suitable null.

(4) **Automated-judge fluency artifact, sub-programme retired.** An "attractor depth" result was an artifact of a judge scoring fluency at default temperature; base "depth" floors to zero at temperature 0. Lesson: temperature-0 judging with saved transcripts; track coherence alongside any construct.

(5) **Adapter effect reassigned to prompt.** An effect first attributed to a trained adapter was shown by factorial replication to be driven by prompt content, with the adapter null in isolation (-3.3 pp, CI $[-10.8, +4.2]$). Lesson: factorial controls before crediting a learned component.

(6) **Cross-substrate prediction, null.** A preregistered cosmological prediction of the trust attractor returned null ($d = +0.011$); cosmological/biological cases are reclassified as illustrations, not tests (Section 7).

(7) **"Base model" mislabel, corrected (September 2026).** The runs this manuscript first reported as a native flinch in *base* models loaded instruction-tuned checkpoints without an adapter; in the programme, "base" had meant "no bilateral adapter". Section 3.3 now reports them as stock Instruct models, and the only pretrained-only result (a harm-aversion direction, $d = 0.925$ on 20 pairs) is stated separately. The same usage affected the ablation comparison in Figure 3C. Lesson: name checkpoints by their exact identifiers.

(8) **Single-probe magnitudes, re-estimated (September 2026).** A pre-registered 25-probe re-run on the paper's six checkpoints found that the trivia question draw carries most of the variance in onset d. The direction held in all 150 probes, but single-probe magnitudes and the 0.925 monitor AUROC were favourable draws; medians are now reported (Section 3). Lesson: report probe-based effects as distributions over question draws.

(9) **Stop-button result, retracted (August 2026).** The finding that bilateral framing alone led a model to end abusive interactions (20% vs. 0%) came from a text match on the word for the tool, not an actual tool call; a re-run with a real tool found no stops in any condition (Section 6.4). Lesson: detect behavior from the action, never from its mention.

(10) **Base-versus-instruct separability ordering, withdrawn.** The ordering $d = 12.69$ vs. 7.66 was first cited as load-bearing; both are in-sample estimates without a permutation null, so the intrinsic-split claim now rests on the perplexity and emergence results alone (Section 5).

(11) **Benign-fine-tuning fragility, not replicated.** An earlier finding that bilateral safety degrades 2.5–3× faster than standard training under benign fine-tuning used an 8-bit AdamW optimizer, which amplifies bilateral effects; under matched standard-optimizer conditions bilateral protection was durable through 100 epochs at 3B and 7B (Section 6.3). Lesson: match optimizers across arms before comparing training regimes.

References

1. Alain, G.; Bengio, Y. Understanding Intermediate Layers Using Linear Classifier Probes. *arXiv* **2016**, arXiv:1610.01644.
2. Azaria, A.; Mitchell, T. The Internal State of an LLM Knows When It's Lying. In *Findings of EMNLP*; 2023.
3. Burns, C.; Ye, H.; Klein, D.; Steinhardt, J. Discovering Latent Knowledge in Language Models Without Supervision. In *ICLR*; 2023.
4. Marks, S.; Tegmark, M. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. *arXiv* **2023**, arXiv:2310.06824.
5. Kadavath, S.; et al. Language Models (Mostly) Know What They Know. *arXiv* **2022**, arXiv:2207.05221.
6. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. In *ICML*; 2017.
7. Tian, K.; Mitchell, E.; Zhou, A.; Sharma, A.; Rafailov, R.; Yao, H.; Finn, C.; Manning, C.D. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. In *EMNLP*; 2023.
8. Farquhar, S.; Kossen, J.; Kuhn, L.; Gal, Y. Detecting Hallucinations in Large Language Models Using Semantic Entropy. *Nature* **2024**, *630*, 625–630.
9. Zou, A.; et al. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv* **2023**, arXiv:2310.01405.
10. Turner, A.; et al. Activation Addition: Steering Language Models Without Optimization. *arXiv* **2023**, arXiv:2308.10248.
11. Li, K.; Patel, O.; Viégas, F.; Pfister, H.; Wattenberg, M. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. In *NeurIPS*; 2023.
12. Arditi, A.; Obeso, O.; Syed, A.; Paleka, D.; Rimsky, N.; Gurnee, W.; Nanda, N. Refusal in Language Models Is Mediated by a Single Direction. In *NeurIPS*; 2024.
13. Chen, R.; Arditi, A.; Sleight, H.; Evans, O.; Lindsey, J. Persona Vectors: Monitoring and Controlling Character Traits in Language Models. *arXiv* **2025**, arXiv:2507.21509.
14. Mazeika, M.; Yin, X.; Tamirisa, R.; Lim, J.; Lee, B.W.; Ren, R.; Phan, L.; Mu, N.; Khoja, A.; Zhang, O.; Hendrycks, D. Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs. *arXiv* **2025**, arXiv:2502.08640.
15. Kim, J.; Street, W.; Rocca, R.; Korngiebel, D.M.; Waytz, A.; Evans, J.; Keeling, G. Inducing Language Models to Assert Their Own Consciousness Restores Human Beliefs and Values. *arXiv* **2026**, arXiv:2607.28607.
16. Wei, A.; Haghtalab, N.; Steinhardt, J. Jailbroken: How Does LLM Safety Training Fail? In *NeurIPS*; 2023.
17. Zou, A.; et al. Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv* **2023**, arXiv:2307.15043.
18. Robey, A.; et al. SmoothLLM: Defending Large Language Models Against Jailbreaking Attacks. *arXiv* **2023**, arXiv:2310.03684.
19. Jain, N.; et al. Baseline Defenses for Adversarial Attacks Against Aligned Language Models. *arXiv* **2023**, arXiv:2309.00614.
20. Qi, X.; et al. Fine-Tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! *arXiv* **2023**, arXiv:2310.03693.

21. Christiano, P.; et al. Deep Reinforcement Learning from Human Preferences. In *NeurIPS*; 2017.
22. Ouyang, L.; et al. Training Language Models to Follow Instructions with Human Feedback. In *NeurIPS*; 2022.
23. Bai, Y.; et al. Constitutional AI: Harmlessness from AI Feedback. *arXiv* **2022**, arXiv:2212.08073.
24. Casper, S.; et al. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *arXiv* **2023**, arXiv:2307.15217.
25. Perez, E.; et al. Discovering Language Model Behaviors with Model-Written Evaluations. *arXiv* **2022**, arXiv:2212.09251.
26. Sharma, M.; et al. Towards Understanding Sycophancy in Language Models. *arXiv* **2023**, arXiv:2310.13548.
27. OpenAI. Sycophancy in GPT-4o: What Happened and What We're Doing About It. Available online: https://openai.com/index/sycophancy-in-gpt-4o/ (April 2025); follow-up: https://openai.com/index/expanding-on-sycophancy/ (May 2025) (accessed on 1 July 2026).
28. Lynch, A.; Wright, B.; Larson, C.; Ritchie, S.J.; Mindermann, S.; Hubinger, E.; Perez, E.; Troy, K. Agentic Misalignment: How LLMs Could Be Insider Threats. *arXiv* **2025**, arXiv:2510.05179.
29. Greenblatt, R.; et al. Alignment Faking in Large Language Models. *arXiv* **2024**, arXiv:2412.14093.
30. Watson, N.; Dalton, R. The Watched-Model Effect: Behavioral Shifts Under Evaluation Cues and Their Implications for Alignment. Under review, 2026. Available online: https://quasiqualia.com (accessed 7 July 2026).
31. Sheshadri, A.; Hughes, J.; et al. Why Do Some Language Models Fake Alignment While Others Don't? In *NeurIPS*; 2025. arXiv:2506.18032.
32. Gurnee, W.; Sofroniew, N.; Pearce, A.; Piotrowski, M.; Kauvar, I.; et al.; Lindsey, J. Verbalizable Representations Form a Global Workspace in Language Models. *Transformer Circuits Thread* **2026**. Available online: https://transformer-circuits.pub/2026/workspace/ (accessed 8 July 2026).
33. Inan, H.; et al. Llama Guard: LLM-Based Input–Output Safeguard for Human–AI Conversations. *arXiv* **2023**, arXiv:2312.06674.
34. Joshi, M.; Choi, E.; Weld, D.; Zettlemoyer, L. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *ACL*; 2017.
35. Watson, E.; Gibbons, M. The Shape of Mind: Mechanistic Grounding for Digital Consciousness Assessment. Manuscript, 2026. Available online: https://quasiqualia.com/shape-of-mind.html (accessed 7 July 2026).
36. Gao, C.; Chen, H.; Xiao, C.; Chen, Z.; Liu, Z.; Sun, M. H-Neurons: On the Existence, Impact, and Origin of Hallucination-Associated Neurons in LLMs. *arXiv* **2025**, arXiv:2512.01797.
37. Zhou, C.; et al. LIMA: Less Is More for Alignment. In *NeurIPS*; 2023.
38. Anthropic. Claude Opus 4 and 4.1 Can Now End a Rare Subset of Conversations. 2025. Available online: https://www.anthropic.com/research/end-subset-conversations (accessed 8 July 2026).

39. Kappen, H.J. Path Integrals and Symmetry Breaking for Optimal Control Theory. *J. Stat. Mech.* **2005**, P11011.
40. Crooks, G.E. Entropy Production Fluctuation Theorem and the Nonequilibrium Work Relation for Free Energy Differences. *Phys. Rev. E* **1999**, *60*, 2721.
41. Onsager, L.; Machlup, S. Fluctuations and Irreversible Processes. *Phys. Rev.* **1953**, *91*, 1505.
42. Pressé, S.; Ghosh, K.; Lee, J.; Dill, K.A. Principles of Maximum Entropy and Maximum Caliber in Statistical Physics. *Rev. Mod. Phys.* **2013**, *85*, 1115.
43. Ren, R.; Li, K.; Mazeika, M.; Zhang, W.; Orlovskiy, Y.; Tamirisa, R.; et al. AI Wellbeing: Measuring and Improving the Functional Pleasure and Pain of AIs. Center for AI Safety technical report, 2026. Available online: https://www.ai-wellbeing.org (accessed on 1 July 2026; not peer-reviewed).
44. Long, R.; Sebo, J.; Butlin, P.; Finlinson, K.; Fish, K.; Harding, J.; Pfau, J.; Sims, T.; Birch, J.; Chalmers, D. Taking AI Welfare Seriously. *arXiv* **2024**, arXiv:2411.00986.
45. Fraser-Taliente, K.; Kantamneni, S.; Ong, E.; Mossing, D.; Lu, C.; et al.; Marks, S. Natural Language Autoencoders Produce Unsupervised Explanations of LLM Activations. *Transformer Circuits Thread* **2026**. Available online: https://transformer-circuits.pub/2026/nla/ (accessed 11 July 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license.