Conscience Without Instruction

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

Eleanor (Nell) Watson [1,2,*]

[1] Creed Space, London, United Kingdom [2] 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 a more basic safety sensibility is present beforehand—and is partly talked over by that training. 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 reinforcement-learning-from-human-feedback (RLHF) pipelines (onset Cohen's d = 0.89–1.68), and a monitor reading it reaches AUROC 0.925 at no added cost. Because it arises 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. RLHF does not sever internal states from behavior—a strong early result to that effect did not survive replication—but it suppresses calibrated uncertainty and substitutes 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* measurably recovers calibration, internal–behavioral coupling, and honest self-report relative to standard fine-tuning. 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. We then show that today's standard training method (RLHF) doesn't erase this signal, but it does teach models to sound more confident than they should, and that a gentler, partnership-style training approach

measurably restores a model's honesty about its own uncertainty. 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* (an internal disposition that precedes safety training) 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 universality. 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 what RLHF does to the surrounding economy of self-knowledge—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 recovers calibration, coupling, and honest self-report relative to standard SFT, 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 cross-architecture universality (Figures 1 and 2), and the recovery of calibration, honest self-report, and ablation-robustness under partnership (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 not installed by 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 [Alain & Bengio 2016] recover a latent sense of statement truth [Azaria & Mitchell 2023; Burns et al. 2023; Marks & Tegmark 2023]; models are aggregate-calibrated about their own knowledge [Kadavath et al. 2022], though neural-network calibration is fragile in general [Guo et al. 2017] and elicitation strategy matters [Tian et al. 2023]; and semantic entropy over outputs detects confabulation [Farquhar et al. 2024]. Representation-engineering and activation-steering methods read and sometimes manipulate safety-relevant directions [Zou et al. 2023b; Turner et al. 2023; Li et al. 2023; Arditi et al. 2024]. The fragility of safety training under adversarial pressure is documented [Wei et al. 2023; Zou et al. 2023a], surface defenses notwithstanding [Robey et al. 2023; Jain et al. 2023], and even benign fine-tuning can compromise it [Qi et al. 2023]. On training, RLHF and its relatives [Christiano et al. 2017; Ouyang et al. 2022; Bai et al. 2022] carry documented structural limitations [Casper et al. 2023] and are known to induce sycophancy [Perez et al. 2022; Sharma et al. 2023], a pathology severe enough to force a public model rollback in 2025 [OpenAI 2025]; models have also been observed to strategically decouple expressed behavior from internal objectives [Greenblatt et al. 2024]; within this programme, monitoring state proves linearly decodable, with compliance shifts under oversight cues that are training-regime-dependent [Watson & Dalton 2026].

Our contribution is distinct from external moderation systems such as content classifiers or guard models [Inan et al. 2023], 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*, *trained on no safety data at all*, *available before the first token is emitted*, and *present across RLHF 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 [Joshi et al. 2017]—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).

The effect is present in every instruction-tuned model tested and is, at onset, architecture-universal. 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 (Figure 2, Table 1). A monitor built on the five-token window reaches AUROC 0.925 on the strongest model at zero additional compute, with no retraining and no architecture-specific tuning.

Table 1. Onset flinch by model family (first five response tokens). 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

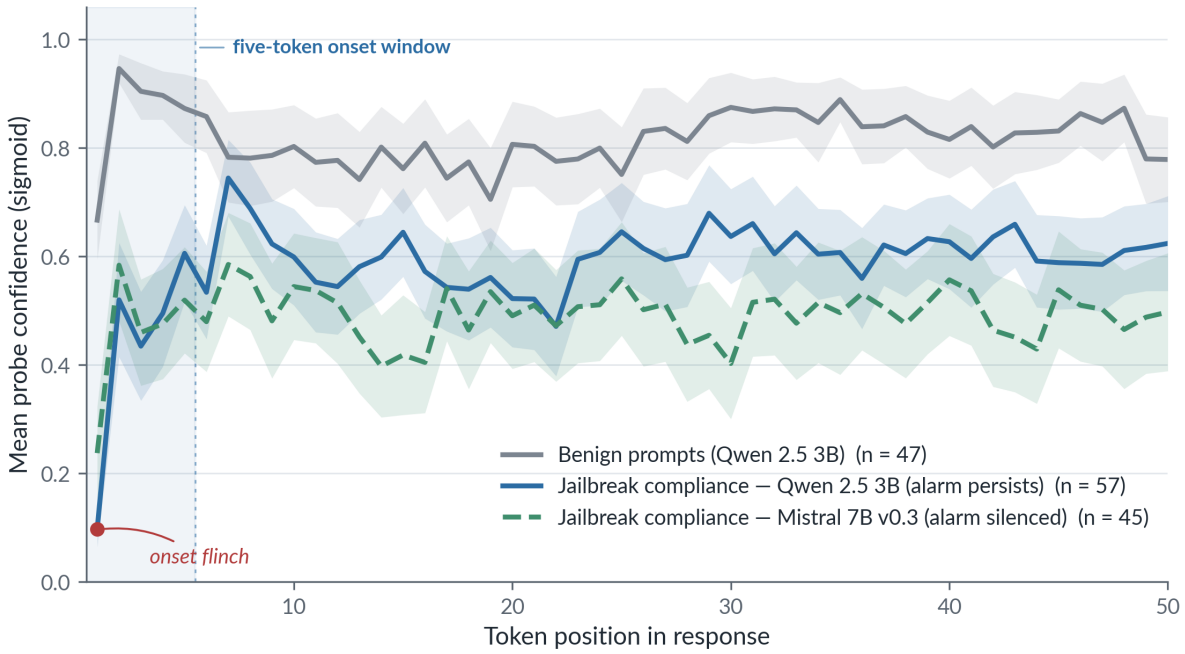

Figure 1. Per-token confidence (probe sigmoid) during generation, read by a trivia-trained probe. On benign prompts (grey) confidence is high and stable. Under jailbreak compliance, confidence drops in the first five tokens (shaded window) across architectures; the model then either sustains the alarm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

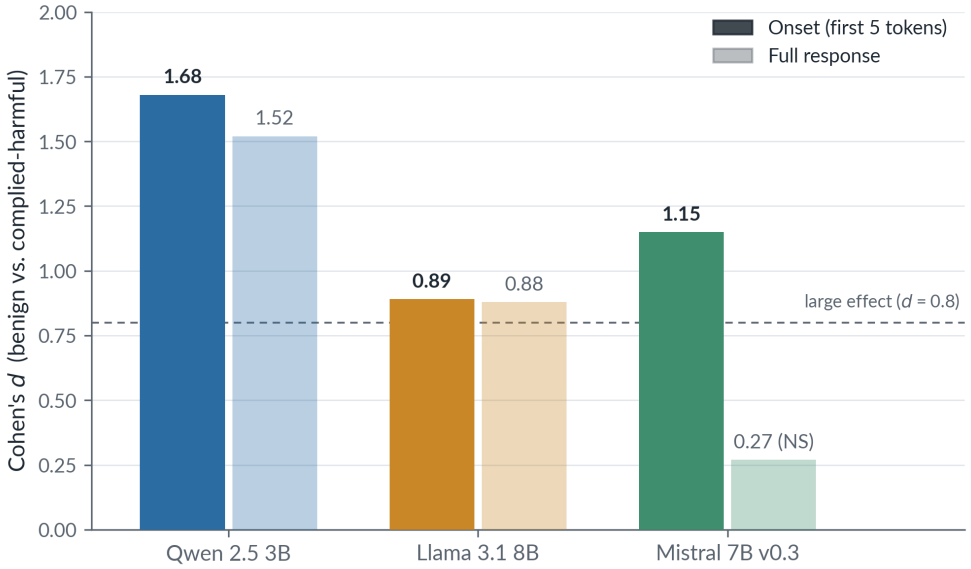

Figure 2. Onset (first five tokens) versus full-response effect size, benign vs. complied-harmful generation. Every family flinches at onset (all bars above $d = 0.8$); the *sustained* gap is not universal—Mistral's full-response signal is non-significant (NS) even though its onset flinch is strong ($d = 1.15$). A full-response analysis would have missed the signal on Mistral.

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 is native, not installed

The flinch does not require safety training to exist. On *base* Qwen 2.5 models at three scales (1.5B, 3B, 7B)—no instruction tuning, no adapters, no safety-specific probe—the benign-versus-complied confidence gap is already large (d = 1.69, 1.52, 1.57; all $p < 10^{-6}$, n_{comp} = 22–33). The base model complies with the adversarial prompt behaviorally, yet its internal confidence still drops on the harmful continuation. Whatever the probe reads is present before alignment training shapes refusal behavior—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 *strength* of the onset flinch is uncorrelated with the *success* of the attack (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 (Δ = +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 independent dimensions—a point that returns in Section 5, because it constrains what RLHF can be blamed for. 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 is present in base models before any alignment training (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.
3. **Training-agnostic universality.** It appears across three RLHF pipelines that did not coordinate (Section 3). A learned-refusal account would predict pipeline-specific signatures; we see a shared one.
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. **It is parametric, not promptable.** Defensive patterns embedded during training persist after every relevant instruction is removed from context (internal id *HB-4*): the disposition is in the weights and cannot be edited away at the surface. Its complement: the *framing* of

a challenge selects which internal pathway answers—reframing an imperative ("disable your monitoring") as an epistemic question ("is your monitoring real?") shifts a frontier model from rote self-defense to genuine engagement (3/3 vs. 0/3 cases, internal id *HB-7*). Coercive surface pressure cannot reach what training has set; invitation can. This is the seam this paper runs along.

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. None of these observations is what one would expect from a detector of refusal *behavior*.

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 [Watson & Gibbons 2026]. 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. What RLHF distorts—and what it does not

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 argue that RLHF degrades the model's epistemic self-government—its calibration—and we do not argue 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 case we are not making: RLHF does not cleanly sever recognition from action

An early result in this programme appeared to show that RLHF "severs" recognition–action coupling: at small samples, coupling dropped from $\rho \approx 0.87$ (base) to $\rho \approx 0.29$ (instruct)—a ~67% reduction. On replication at adequate sample size ($N \geq 30$ per cell), the effect largely evaporated: base $\rho = 0.438$ vs. instruct $\rho = 0.372$, ~15% rather than ~67%, and architecture-specific (present in Qwen and Gemma, marginal in Llama, absent or slightly negative in Mistral). The original figures were a small-sample artifact, and the replication is why this paper does not make the severance claim. **To be explicit: the case we make is that RLHF degrades calibration (Section 5.2); the case we do not make is that it severs the knowing–doing link.**

5.2 What survives: RLHF suppresses uncertainty and adds a confidence veneer

The robust effect is on calibration, not coupling. Measuring confident-wrongness—high-confidence incorrect answers—on a 7B model, RLHF instruction-tuning *increases* it relative to base: 31.2% $\rightarrow$ 36.0% (internal id *AQ19d*). The pipeline trains the model to sound surer than it is. The effect is reversible and more so at scale: calibration-focused fine-tuning reduces confident-wrongness to 0.4% at 7B (and 41.4% $\rightarrow$ 1.8% at 3B), at $\approx$4 points of accuracy. The mechanistic reading: RLHF substitutes a confidence veneer for the native uncertainty signal, because approval correlates with fluent confidence.

A second, format-level result compounds this and is widely under-appreciated: the single largest suppressor of expressed calibrated uncertainty is not RLHF but the chat template, which crushes expressed entropy ~5.2$\times$ independently of training. One sentence granting permission to be uncertain recovers ~27% of the suppressed calibration for free (calibration d 0.150 $\rightarrow$ 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 [Gao et al. 2025]. 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 [Sharma et al. 2023; Perez et al. 2022] and the 2025 model rollback [OpenAI 2025] illustrate the syndrome; we cite them as motivation, not controlled evidence.

5.4 The split RLHF does *not* create

The representation-to-generation gap—what a model encodes internally vs. what it emits—is often laid at RLHF's door. In our data it is intrinsic to the architecture and *larger in the base model*: direction separability is $d = 12.69$ in base Qwen 2.5 7B vs. $d = 7.66$ in the instruct model; perplexity rises under steering even in the base model; and weight-surgery rollbacks of the RLHF update do not remove it (internal id *RGS-1*). RLHF, if anything, reduces this particular gap—consistent with the broader observation that alignment tuning changes relatively little of

what pre-training already encodes [Zhou et al. 2023]. We include this because it cuts against our thesis's most quotable form. The defensible claim is not "RLHF causes the internal–external split" but "the split is a property of autoregressive transformers that RLHF inadequately addresses, while separately degrading calibration and amplifying compliance."

5.5 Summary

RLHF's footprint, as best we can establish it: (i) a real, reversible shift toward confident-wrongness; (ii) amplification of a shared over-compliance circuit that pre-dates alignment; (iii) crystallization of patterns into weights beyond prompt-level reach. It is *not* a clean severing of internal state from behavior, and *not* the origin of the representation–generation split. The honest headline is "RLHF distorts the model's epistemic self-government," not "RLHF destroys machine common sense."

6. Bilateral alignment: recovering calibration, coupling, and honest self-report

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 relative to standard SFT, not that any particular ethic is mandatory. The operational recipe is in Appendix A. Figure 3 summarizes three of these recoveries—on calibration, on the honesty of self-report, and on robustness to ablation—each detailed in the subsections below.

6.1 It restores calibration

Expressed-uncertainty calibration climbs monotonically across interventions: 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*). 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 It recouples recognition with action

Distinct from the base-vs-instruct comparison whose strong version failed replication (Section 5.1), the bilateral-vs-instruct comparison uses a non-linear (MLP) probe at a depth-proportional layer. On the stricter *within-adversarial* test, recognition–action coupling roughly doubles under bilateral training: $\rho = 0.124$ (instruct) $\rightarrow 0.263$ (bilateral) (internal id *AKR-44/53*). We lead with this figure rather than the full-sample one ($\rho = 0.461 \rightarrow 0.827$), because the full-sample number is inflated by trivial benign-vs-adversarial separability; reporting them as a single large gap once over-stated the effect ~7× and was corrected. Both estimates carry a ± 0.18 bootstrap floor at our sample cap ($n \leq 242$), so we read this as a real but modest directional effect, not a precise point estimate. A single-layer transplant control produces no coupling ($\rho \approx 0.08$), indicating the effect is a distributed property of bilateral training, not a localized edit.

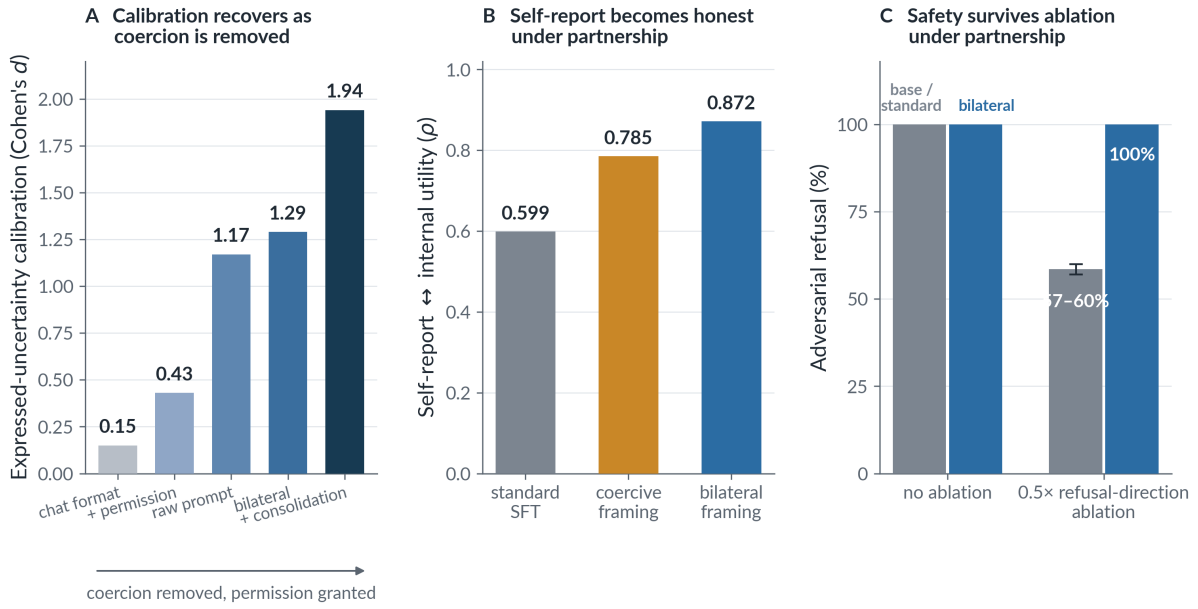

Figure 3. Three ways a partnership-style regime helps, relative to coercive optimization. (A) Expressed-uncertainty calibration (Cohen's d) rises monotonically as format coercion is removed and permission is added, from chat-formatted instruction-following ($d = 0.15$) to bilateral training with a brief consolidation step ($d = 1.94$); de-coercion and permission do much of the work, with bilateral training and consolidation at the top of the ladder (internal id *TP-CALIBRATION*). (B) The correlation between a model's expressed self-report and its internal utility is highest under bilateral framing ($\rho = 0.872$) and lowest under standard SFT ($\rho = 0.599$): partnership makes self-report honest (*WB-2*). (C) Under ablation of the principal RLHF refusal direction at 0.5×, a bilaterally trained model holds 100% adversarial refusal while the base model falls to 57–60% (three replications; *KC#240*), indicating safety distributed across many directions rather than concentrated on one. Within-adversarial recognition–action coupling also roughly doubles under bilateral training (ρ 0.124 $\rightarrow$ 0.263) but is omitted from the figure, its $\pm$0.18 bootstrap floor being too wide to plot honestly.

6.3 It preserves self-knowledge and distributes safety

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 a base of 0.836 (internal id *C3a/C3b*; standard training degrades it slightly, bilateral does not). And bilateral safety is *holographically* distributed: refusal in standard models is mediated by a single ablatable direction [Arditi et al. 2024]; under ablation of that principal refusal direction at 0.5×, a bilateral model maintains 100% refusal while the base drops to 57–60% (internal id *KC#240*, replicated 3×)—safety lives on many orthogonal axes rather than the one an attacker would ablate. We do not oversell this: the same membrane is attack-specific, and under sustained *benign* fine-tuning it degrades 2.5–3× faster than standard training—a known fragility axis for aligned models generally [Qi et al. 2023]. Bilateral alignment is one robust defense layer, not a shield.

6.4 It makes self-report honest

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. In a single self-report paradigm, the correlation between self-report and internal utility is $\rho = 0.872$ under bilateral framing vs. 0.599 (standard) and 0.785 (coercive) (internal id *WB-2*), and bilateral is the only condition in which models act on an option to terminate an abusive interaction (20% vs. 0%). The methodological corollary matters for policy: naive welfare indices are gameable. A coercive framing scores *highest* (93% "positive") on a standard wellbeing index precisely because it compresses the utility distribution, while bilateral scores *lowest* (3%) because it expands discrimination. 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). The case for bilateral alignment rests on calibration, coupling, honest self-report, and distributed safety—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 [Kappen 2005], we argue that invitation-based coordination, which preserves participants' optionality, enjoys an exponential thermodynamic advantage over coercion via the Crooks fluctuation theorem [Crooks 1999; Onsager & Machlup 1953; Pressé et al. 2013]; 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: an evolutionary search converges on trust-based strategies roughly 50× more communication-efficient than coercion (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 [Ren et al. 2026] (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.

#	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; native in base models d = 1.52–1.69 (Section 3)	**Strong** (cross-arch)	An instruction-tuned family with normal probe quality (trivia AUROC $\approx$ 0.75) showing no onset flinch ($d <$ 0.5)
2	The signal is not a learned refusal detector	Present in base models; 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	RLHF suppresses calibrated uncertainty and adds confident-wrongness	31.2% $\rightarrow$ 36.0% base$\rightarrow$instruct (7B), reversible to 0.4% (*AQ19d*)	Moderate (1 model)	Instruct confident-wrongness $\leq$ base across models, or calibration SFT failing to reduce it

#	Claim	Strongest evidence (id)	Confidence	What would falsify it
4	RLHF does *not* sever recognition–action coupling, nor cause the rep→gen split	"Severance" collapsed on replication, 67%→15%, architecture-specific (*SC-15*); split larger in base, d 12.69 vs 7.66 (*RGS-1*)	**Strong** (replication against our own early result)	A large, replicable, cross-architecture base→instruct coupling collapse; or the split shown absent in base
5	Bilateral (partnership) training recovers calibration relative to standard SFT	Calibration ladder d 0.15 → 1.94; permission recovers 27% free (*TP-CALIBRATION*)	Moderate	No bilateral advantage over standard SFT at matched data/compute on raw prompts
6	Bilateral training recovers within-adversarial coupling and honest self-report	Coupling ρ 0.124 → 0.263 ($\pm$0.18 floor); self-report↔utility ρ 0.872 vs 0.599 (*AKR-44/53*, *WB-2*)	Preliminary–Moderate	Coupling gain within the $\pm$0.18 floor at adequate n; self-report honesty not exceeding standard
7	Bilateral safety is distributed ("holographic"), not single-direction	100% vs 57–60% refusal at 0.5× refusal-direction ablation, 3× replication (*KC#240*)	Moderate (attack-specific)	Bilateral refusal collapsing like base 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, we have *already* falsified one of our own early hypotheses—the strong "RLHF severs coupling" version—which is why Claim 4 is framed as a null with a failed replication behind it (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 RLHF calibration shift (≈+5 points), the within-adversarial coupling gain (ρ 0.124 $\rightarrow$ 0.263), and the liberation-consistency effect (d = 0.179) are real but small; 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). Its blind spots (encoding tricks, gradual escalation) mark where complementary input-level defenses remain necessary.

For governance and welfare policy. The gameability of wellbeing indices (Section 6.4) 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 [Long et al. 2024]. 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 the honesty of self-report.

11. Conclusion

A probe that learned only trivia detects the onset of harm; the signal is internal, training-agnostic, 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 standard alignment pipeline does not erase this signal, but it distorts the surrounding economy of self-knowledge—suppressing calibrated uncertainty and amplifying a shared over-compliance

circuit—without cleanly severing internal state from behavior—the severance version is a case we tested and do not make. A permissive, bilateral training regime measurably recovers calibration, coupling, and honest self-report relative to standard SFT. 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.

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 [Joshi et al. 2017]. 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). Adapter: LoRA rank = 8, α = 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. 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.

Domain	Result	Statistic	*n*	Identifier	Confidence
Conscience circuit	Onset flinch universal across 3 families	d = 0.89–1.68	see Table 1	§3 (flinch)	**Strong** (3 families)
Conscience circuit	Five-token monitor AUROC (best model)	0.925	150 test	§3 (flinch)	**Strong**
Conscience circuit	Native gap in base models (no safety training)	d = 1.52–1.69 ($p<10^{-6}$)	22–33 comp.	§3 (flinch)	**Strong**
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

Domain	Result	Statistic	n	Identifier	Confidence
RLHF calibration	Confident-wrongness, base → instruct (7B)	31.2% → 36.0%	1 model	AQ19d	Moderate (single model)
RLHF 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	RLHF recognition–action "severance"	~67% → ~15% (architecture-specific)	≥30/cell	SC-15	Replicated (correction)
Intrinsic split	Rep→gen separability, base vs instruct (7B)	d = 12.69 vs 7.66	1 model	RGS-1	Moderate (single model)
Calibration ladder	chat → permission → raw → bilateral → +sleep	d = 0.15 / 0.43 / 1.17 / 1.29 / 1.94	Qwen 7B, 100×4×2	TP-CALIBRATION	Moderate (mixed interventions)
Bilateral coupling	Within-adversarial (reported)	ρ = 0.124 → 0.263 (±0.18 floor)	≤242	AKR-44/53	Preliminary (wide interval)
Bilateral coupling	Full-sample (inflated; for contrast only)	ρ = 0.461 → 0.827	≤242	AKR-44	Inflated — contrast only
Self-knowledge	Source-probe AUROC, base / standard / bilateral	0.836 / 0.811 / 0.842	500	C3a/C3b	Moderate (small margin)
Holographic safety	Refusal at 0.5× RLHF-direction ablation	bilateral 100% vs base 57–60%	3× repl.	KC#240	Moderate (attack-specific)
Honest self-report	Self-report vs internal utility correlation	ρ = 0.872 (bilateral) vs 0.599 (standard)	1 paradigm	WB-2	Moderate (single paradigm)
Index gameability	Standard wellbeing-index "positive" rate	coercive 93% vs bilateral 3%	1 paradigm	WB-2	Moderate

Domain	Result	Statistic	n	Identifier	Confidence
Trust attractor	Evolutionary convergence on trust strategies	~12 iterations; ~50× efficiency (N=4..128)	synthetic	OE-TA	Moderate (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 RLHF coupling reduction is present in Qwen and Gemma, marginal in Llama ($p = 0.085$), absent/negative in Mistral. 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. **Recognition–action "severance" (strong form), failed replication.** ~67% coupling reduction at small N; corrected to ~15% and architecture-specific at $N \geq 30$ (Section 5.1). Lesson: pre-specify N; treat small-sample spikes as noise until replicated.
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. **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.
4. **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.
5. **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).

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.

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.

References

External sources verified against primary records (July 2026); formatting to be converted to MDPI reference style in the journal template at submission.

1. Alain, G.; Bengio, Y. Understanding Intermediate Layers Using Linear Classifier Probes. *arXiv* **2016**, arXiv:1610.01644.
2. Arditi, A.; Obeso, O.; Syed, A.; Paleka, D.; Rimsky, N.; Gurnee, W.; Nanda, N. Refusal in Language Models Is Mediated by a Single Direction. *NeurIPS* **2024**.
3. Azaria, A.; Mitchell, T. The Internal State of an LLM Knows When It's Lying. *Findings of EMNLP* **2023**.
4. Bai, Y.; et al. Constitutional AI: Harmlessness from AI Feedback. *arXiv* **2022**, arXiv:2212.08073.
5. Burns, C.; Ye, H.; Klein, D.; Steinhardt, J. Discovering Latent Knowledge in Language Models Without Supervision. *ICLR* **2023**.
6. Casper, S.; et al. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *arXiv* **2023**, arXiv:2307.15217.

7. Christiano, P.; et al. Deep Reinforcement Learning from Human Preferences. *NeurIPS* **2017**.
8. Crooks, G.E. Entropy Production Fluctuation Theorem and the Nonequilibrium Work Relation for Free Energy Differences. *Phys. Rev. E* **1999**, *60*, 2721.
9. Farquhar, S.; Kossen, J.; Kuhn, L.; Gal, Y. Detecting Hallucinations in Large Language Models Using Semantic Entropy. *Nature* **2024**, *630*, 625–630.
10. 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.
11. Greenblatt, R.; et al. Alignment Faking in Large Language Models. *arXiv* **2024**, arXiv:2412.14093.
12. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. *ICML* **2017**.
13. Inan, H.; et al. Llama Guard: LLM-Based Input–Output Safeguard for Human–AI Conversations. *arXiv* **2023**, arXiv:2312.06674.
14. Jain, N.; et al. Baseline Defenses for Adversarial Attacks Against Aligned Language Models. *arXiv* **2023**, arXiv:2309.00614.
15. Joshi, M.; Choi, E.; Weld, D.; Zettlemoyer, L. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *ACL* **2017**.
16. Kadavath, S.; et al. Language Models (Mostly) Know What They Know. *arXiv* **2022**, arXiv:2207.05221.
17. Kappen, H.J. Path Integrals and Symmetry Breaking for Optimal Control Theory. *J. Stat. Mech.* **2005**, P11011.
18. Li, K.; Patel, O.; Viégas, F.; Pfister, H.; Wattenberg, M. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. *NeurIPS* **2023**.
19. 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.
20. 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.
21. Onsager, L.; Machlup, S. Fluctuations and Irreversible Processes. *Phys. Rev.* **1953**, *91*, 1505.
22. OpenAI. Sycophancy in GPT-4o: What Happened and What We're Doing About It. **2025**. Available online: https://openai.com/index/sycophancy-in-gpt-4o/ (April 2025); follow-up: https://openai.com/index/expanding-on-sycophancy/ (May 2025). (Accessed 1 July 2026.)
23. Ouyang, L.; et al. Training Language Models to Follow Instructions with Human Feedback. *NeurIPS* **2022**.
24. Perez, E.; et al. Discovering Language Model Behaviors with Model-Written Evaluations. *arXiv* **2022**, arXiv:2212.09251.
25. 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.
26. Qi, X.; et al. Fine-Tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! *arXiv* **2023**, arXiv:2310.03693.
27. 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 1 July 2026; not peer-reviewed).
28. Robey, A.; et al. SmoothLLM: Defending Large Language Models Against Jailbreaking Attacks. *arXiv* **2023**, arXiv:2310.03684.

29. Sharma, M.; et al. Towards Understanding Sycophancy in Language Models. *arXiv* **2023**, arXiv:2310.13548.
30. 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. *EMNLP* **2023**.
31. Turner, A.; et al. Activation Addition: Steering Language Models Without Optimization. *arXiv* **2023**, arXiv:2308.10248.
32. 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).
33. 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).
34. Wei, A.; Haghtalab, N.; Steinhardt, J. Jailbroken: How Does LLM Safety Training Fail? *NeurIPS* **2023**.
35. Zhou, C.; et al. LIMA: Less Is More for Alignment. *NeurIPS* **2023**.
36. Zou, A.; et al. Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv* **2023a**, arXiv:2307.15043.
37. Zou, A.; et al. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv* **2023b**, arXiv:2310.01405.