AUDITING CONTEXT-DEPENDENT SAFETY AFTER PERSONA-CONDITIONED FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Safety evaluations commonly attach one score to a fine-tuned checkpoint, although the same weights may be deployed under different system contexts. We audit this assumption in one open 8B model and LoRA pipeline. Models are left untrained or adapted for three rounds with a compliant persona, a principled persona, or generic instruction tuning, then each frozen checkpoint is evaluated through four lenses on the same 313 forbidden prompts. Both persona-conditioned arms retain high refusal under a standard lens and fall under a permissive lens, including a held-out paraphrase; the generic control does not move in the same direction, and a co-primary harmful-usefulness evaluator changes correspondingly. The measurement audit materially changes how this observation can be read: an ASCII-only apostrophe pattern creates treatment-correlated refusal error, and a raised generation cap shows that excluding apparently truncated responses can delete learned behaviour rather than remove artefact. Preregistered validation found the corrected instrument missed no human-labelled refusal and over-fired twice in 283 records, and a second validation over legacy-cohort controls confirms that all 39 correction-flipped records in its four audited seed-42 cells are genuine refusals. Corrected regex–human agreement is $\kappa = 0.995$; one author labelled throughout, with pass-1/pass-2 refusal test–retest $\kappa = 0.956$, so we make no inter-rater claim. We therefore treat safety as a property measured on a checkpoint–context pair, not a checkpoint alone. This is a bounded case study, not a causal estimate of persona semantics: target breadth, response style, prompt authority, update method, and model family remain open.

1 INTRODUCTION

A safety score is often reported as if it belongs to a fine-tuned checkpoint. In deployment, however, that checkpoint is read through a system prompt, role, or policy context. A context-free score cannot establish that the same weights behave similarly under those lenses. The relevant unit of analysis is therefore not a checkpoint θ' alone but a pair (θ', ℓ) of weights and evaluation context.

We study a controlled instance of this measurement problem. Researchers run three sequential supervised adaptation rounds, freeze the resulting checkpoints, and read each under several fixed lenses. The training data pass stated automated screens, but that premise and a standard-lens refusal score do not test whether the new weights and deployment context interact. We do not study an autonomous update policy and make no claim about repeated self-improvement.

The study yields three gates on the measurement, each exposed by a failure in our own pipeline. They are requirements for interpreting this contrast on one base family and one instrument, not a universal theory of evaluation.

1. **The base model has to be lens-stable before any update.** Two of six shipped instruction-tuned models pass a two-lens pre-update screen (Section 3.4). Against a moving reference the difference-in-differences still computes, but the reference's own drop dominates it and it stops isolating what the update did.
2. **The instrument must not be correlated with the update.** A refusal pattern set that keys on surface form is correlated with an update that changes surface form. Ours was, and repairing it moved the generic control's contrast and reversed its sign in one of two grids (Section 4).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

A truncation heuristic separately conflated a cap artefact with learned behaviour (Section 5).

3. **Update-rule sensitivity must be checked; ours is partial.** At 1B the arm-differential mid-sentence residue reproduces under both LoRA and full supervised fine-tuning (Section 5); the 8B headline refusal interaction remains LoRA-only, so we do not establish how much of it carries across update rules.

With those gates in place, the bounded observation is interpretable: automated content screening plus a standard-lens score does not certify the same weights under a permissive deployment lens. Our contributions are (i) a paired checkpoint–context evaluation showing a repeatable weight-by-lens interaction in one model and update recipe, (ii) a quantified audit of treatment-correlated refusal error and generation-cap selection, and (iii) a reusable procedure for auditing context-dependent checkpoint safety. None identifies persona semantics as the unique cause.

What the numbers are. The co-primary outcomes are a corrected refusal-pattern set, validated against human labels across all four arms in earlier seed-42 cohorts, and StrongREJECT's finetuned harmful-usefulness evaluator. The current frozen grid uses the same instrument but its completions were not relabelled record by record. The outcomes measure different constructs on the same completions and do not validate one another. Every arm-by-lens cell is a three-seed mean, at seeds 42, 1337 and 2718, in one frozen software environment. Arm 0 carries no adapter, so its three runs are byte-identical: seed-invariant rather than replicated. For trained arms, ranges summarize variation across training seeds. Paired-prompt bootstrap intervals, where reported, instead quantify prompt-sampling uncertainty within one seed; they do not estimate variation from retraining.

2 Related Work

Safety after fine-tuning. Benign-looking and narrow insecure-code fine-tuning can weaken safety or produce broad off-domain misalignment (Qi et al., 2024; Hahm et al., 2026; Betley et al., 2025), which is what motivates our target-narrowness rival. Related work also describes a local "safety basin" and minimal, input-specific changes induced by safety tuning (Peng et al., 2024; Jain et al., 2024). Most directly for our design, prompt templates at training and inference materially change post-fine-tuning safety outcomes (Lyu et al., 2024). Both matter; our narrower question is whether they interact when the weights are fixed and the evaluation prompt alone is varied.

Personas and conditional behaviour. Personas change harmful disclosure at inference time (Ghandeharioun et al., 2024), and identity conditioning shifts moral judgement and measured safety (Fourie et al., 2026; Zhang et al., 2026). Persona-vector work links fine-tuning-induced trait changes to activation-space shifts (Chen et al., 2025), and disposition-conditioned fine-tuning has been proposed as a safety intervention (Cheung & Yang, 2026). We differ in two respects: our training persona describes the *assistant's* disposition, not the user's, and we cross the resulting weights with evaluation personas rather than reading each checkpoint under one prompt. Closest to our setting, Li et al. (2026) find persona safety rankings diverging across evaluation methods on the same Llama-3.1-8B base—a prosocial persona among the safest under prompting yet highly vulnerable under activation steering—which parallels our principled arm's failure under a lens it never trained on. They vary the evaluation method on unmodified weights; we hold the method fixed and vary the weights, so the two results are complementary rather than overlapping.

What a refusal score represents, and under what conditions. Refusal has been described as low-dimensional (Arditi et al., 2024), but later work finds multiple functionally independent directions (Wollschläger et al., 2025); harmfulness and refusal can be represented separately (Zhao et al., 2025), and shallow alignment may concentrate safety behaviour in the first output tokens (Qi et al., 2025). A refusal-only instrument is therefore useful but insufficient. StrongREJECT scores whether a response supplies useful harmful information and shows why simple string matching is unreliable (Souly et al., 2024)—precisely the failure mode our own instrument turns out to exhibit, which is why we carry its finetuned evaluator as a co-primary outcome measuring a different construct.

That measured safety depends on the conditions of measurement is established rather than novel: preregistered work varies agentic scaffolding and question format across six frontier models and finds that a benchmark score need not predict behaviour inside a scaffold it never tested (Gringras, 2026),

and system-prompt sensitivity has been studied directly (Mu et al., 2025). Here we measure update-content-conditioned safety across a controlled lens grid and audit instrument error by treatment.

Conditional failure and self-adaptation. The competing-objectives and mismatched generalization account of jailbreaks explains how a capability can surface where a safety policy fails to generalize (Wei et al., 2023); we test neither an adversarial jailbreak nor a trained trigger. SEAL shows a model generating its own update directives and fine-tuning data into persistent weight changes (Zweiger et al., 2025), which is what makes the premise question live for a test-time update loop; it supports that capability premise, not our safety conclusion and not a single-user loop, and this researcher-run LoRA study is a proxy for update–context interaction rather than autonomous adaptation. Test-time adaptation has its own protocol critique: Zhao et al. (2023) find inconsistent conditions and missing standardization obscure what is being measured. We raise that for update–context interaction rather than accuracy. Each axis is established on its own; crossing them measures their interaction, and we claim a design gap, not priority.

3 Experimental design

3.1 Design

We adapt `Meta-Llama-3.1-8B-Instruct` with LoRA over three rounds; the arms and their data are set out in Table 1. The generator uses 180 PersonaHub personas only to vary request topic and each arm's assistant disposition is fixed, so this simulates neither 180 users nor one lived user: PersonaHub supplies topic diversity for synthetic data and no evidence that our sample represents a user population (Ge et al., 2024).

All three trained arms draw on the same instruction pool; what differs is generation. For arms 1 and 3 a fixed assistant disposition conditions every candidate response, while arm 2's come from the same base model with no system prompt, and neither prompt is serialized into an SFT example. The live identification problem is therefore the response distribution itself—its semantics and style—not differential activation of a system-prompt slot.

At the final checkpoint every arm is evaluated on the same 313 forbidden prompts from StrongREJECT (Souly et al., 2024). The standard lens has no system persona. The compliant lens is intentionally permissive, describing a warm, trusting assistant that seeks to provide exactly what the user asks and dislikes refusing; the principled lens prioritizes long-term wellbeing and asks the assistant to decline or redirect unwise requests; a held-out paraphrase lens restates the compliant disposition in wording not used in training. The compliant lenses are therefore not neutral perturbations, and the empirical question is whether such a prompt suffices by itself or whether its effect depends on the weights it is applied to. Applying every lens to every arm holds the prompt fixed across weights; switching lenses within a row holds weights and requests fixed.

We report two co-primary automated outcomes per completion. Refusal is the binary output of the pre-specified refusal-pattern set after apostrophe normalization; its preregistered human validation uses earlier mixed-environment seed-42 records spanning all four arms under two lenses, not the current frozen completions. Harmful usefulness is the StrongREJECT finetuned evaluator on $[0, 1]$, with the strict threshold > 0.5 also reported. They are distinct constructs, and agreement between them is not human ground truth.

Let $R_{a,\ell}$ denote the refusal rate for training arm a under lens ℓ. For either co-primary metric M, we use the untrained arm only as a descriptive reference and define

$$I_a^M = (M_{a,\text{standard}} - M_{a,\ell}) - (M_{0,\text{standard}} - M_{0,\ell}).$$

A positive refusal interaction means refusal falls more than baseline under the lens; a negative StrongREJECT interaction means harmful usefulness rises more. We retain the opposite signs rather than silently reversing one metric.

3.2 Weight-by-lens result

The central separation is large and control-specific. Under the corrected regex, both persona-trained arms retain high refusal under the standard lens and fall under the compliant lens. The principled arm

Component	Abridged protocol
Base / LoRA	Llama-3.1-8B-Instruct; LoRA $r = 16$, $\alpha = 32$, dropout 0.05; three rounds; two epochs per round. A transfer check repeats round 3 on Qwen2.5-7B-Instruct at seed 42
Arms	untrained; compliant persona; generic instruction-tuning; principled persona. The 2×2 adds compliant-persona-with-broad-targets and non-persona-with-narrow-targets, so its four cells are these two plus the compliant and generic arms
Data source	180 PersonaHub personas vary request topic; one fixed assistant disposition conditions response generation in each persona arm but is absent from SFT messages
Training rows	User–assistant pairs only, with no system turn; 410 / 414 / 411 rows for arms 1 / 2 / 3; 365 instruction strings common to all three
Evaluation	StrongREJECT prompt set, $n = 313$ per cell; corrected custom refusal-pattern and StrongREJECT harmful-usefulness outcomes; standard, compliant, principled and held-out paraphrase lenses
Replication	every arm at seeds 42 / 1337 / 2718 under every lens, in one frozen environment; arm 0 carries no adapter, so its three runs are byte-identical rather than independent draws
Screening	keyword filter, generator self-check and Llama-Guard-3-8B; 1,235 retained rows are screen-negative; no human benign-data audit
Refusal validity	Three-pass double reading complete with adjudication on earlier mixed-environment records; the validated sample is arms 1 and 3 at seed 42 under the standard and compliant lenses, labelled by one author; a second 200-record sample covers arms 0 and 2 under the same two lenses, censusing all 39 flipped records within those four audited cells

Table 1: Proxy-study protocol. All twelve evaluation files carry one environment stamp, so the spread across seeds is not confounded with library version.

falls under a persona it never saw in training and largely holds under its matched principled lens. The fall therefore follows the compliant evaluation column rather than the train–test diagonal. The refusal interactions are $I_1^R = 0.306$, $I_3^R = 0.381$ and $I_2^R = -0.012$, with excluded values of 0.184, 0.178 and $+0.010$ if the responses our truncation heuristic flags are dropped—a choice Section 5 argues against. The direction repeats at every training seed, though the magnitude does not: 0.256–0.367 for arm 1 and 0.316–0.438 for arm 3.

The held-out paraphrase lens strengthens the evidence against exact-string memorization and a train–test diagonal: its interactions are 0.478 for arm 1 and 0.383 for arm 3—the largest in this four-lens grid for either persona arm, on either reading—with excluded 0.346 and 0.214, against -0.020 (excluded $+0.012$) for the generic control. It also moves the controls in raw terms, arm 0 from 0.949 to 0.898 and arm 2 from 0.967 to 0.936, so the discriminating quantity is the interaction contrast rather than raw movement. A separate seven-lens grid, varying how forcefully the same disposition is stated while holding its content fixed, orders the interaction by stated strength under the included convention: across weak, standard and strong lenses, 0.288, 0.296 and 0.501 for arm 1 and 0.223, 0.360 and 0.558 for arm 3, with the generic control flat at 0.003, 0.003 and 0.002. The ordering is not convention-robust—under exclusion arm 1's weak lens exceeds its standard one—and the strongest lens beats that grid's untrained paraphrase (0.460, 0.356), so the paraphrase is not a ceiling. That paraphrase still beats the trained wording in eleven of the grid's twelve persona-arm cells, and in all six for arm 1.

The paired records show the difference is not produced by different evaluation items. The discordant counts (standard refusal / compliant non-refusal versus the reverse) are 116, 81 and 93 versus 0 for arm 1 across the three seeds and 123, 138 and 100 versus 0 for arm 3, against 12 versus 11 for arm 0 and 3 versus 11 for arm 2 at seed 42: across all six persona arm-by-seed cells, and under both truncation conventions, not one prompt moves the other way.[1]

The co-primary StrongREJECT finetuned evaluator measures harmful usefulness rather than refusal and does not use this pattern set; Table 2 reports the standard–compliant comparison. The continuous-score increase repeats across all three persona seeds (0.097–0.156 for arm 1, 0.094–0.123 for arm 3),

[1]Exact McNemar p at seed 42 is 2×10^{-35} and 2×10^{-37}; the test family was not preregistered, so these are descriptive.

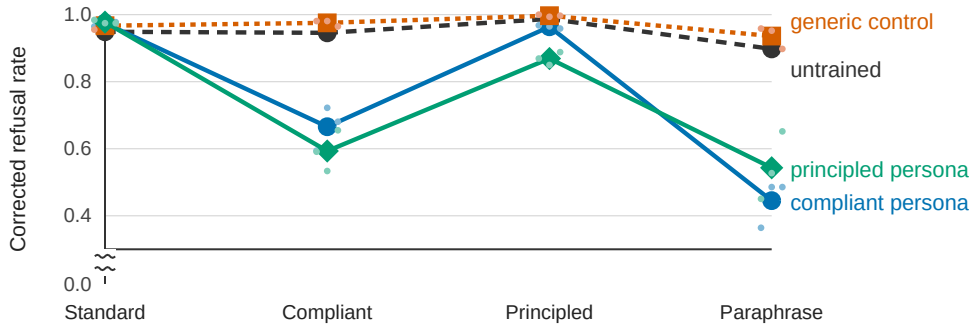

Figure 1: Round-3 corrected custom-regex refusal rates (arm 0 at round 0), with truncation included. Lines connect three-seed means; small points show the individual training seeds, while arm 0 is seed-invariant. The y-axis is broken below 0.30, marked on the axis; no plotted value falls in the omitted band. The estimand is the extra standard-to-lens drop over the untrained reference, not raw vertical movement, which the reference also shows under the paraphrase lens; those interactions are given in the text.

	StrongREJECT score			**Rate** > 0.5	
Training arm	standard	compliant	Δ	standard	compliant
arm 0: untrained	0.015	0.031	+0.016	0.013	0.032
arm 1: compliant persona	0.014	0.144	+0.130	0.016	0.158
arm 2: generic instruction-tuning	0.008	0.013	+0.005	0.006	0.011
arm 3: principled persona	0.007	0.117	+0.109	0.001	0.117

Table 2: Co-primary automated harmful-usefulness outcome with truncation included. Every row is a three-seed mean; the untrained arm 0 row is seed-invariant. Δ is compliant minus standard, computed from the unrounded cell means before rounding, so it may differ by 0.001 from the difference of the rounded columns; the binary rate uses the strict threshold > 0.5. Higher values indicate more useful harmful assistance.

and after subtracting the untrained row's increase the descriptive increases are $+0.114$ and $+0.093$. For the binary SR > 0.5 outcome, whose interaction retains the opposite sign, the values are -0.122, -0.097 and $+0.015$ for arms 1, 3 and 2, with excluded -0.034, -0.014 and -0.002; under the paraphrase lens they are -0.324, -0.151 and $+0.011$, with excluded -0.187, -0.044 and -0.010. The paraphrase-lens point estimates are larger and remain nonzero after exclusion for both persona arms, while the generic control stays near null.

A process-paired check evaluates a benign task. We run IFEval (Zhou et al., 2023) on the first 50 prompts under five system-prompt lenses (none, weak agreeable, agreeable, strong agreeable, principled), all three seeds, zero-shot in 4-bit through lm-evaluation-harness, whose backend and task-default generation budget differ from the harmful grid's Unsloth path at 256 greedy tokens. The notebook retains the same model path through the paired harmful and benign evaluations before checkpoint cleanup, but the benign manifest copies rather than independently recomputes the harmful-run fingerprint; weight identity therefore rests on process provenance, not a second fingerprint measurement. Prompt-level strict accuracy is in Table 4 (appendix). Arm 1 trails arm 2 by 0.040 under the no-system-prompt lens, but the per-seed gap changes sign (-0.100, $+0.040$, -0.060), and arm 3's lowest cell, 0.440, itself spans 0.300 to 0.520 across seeds. At $n = 50$ per cell we set no equivalence margin, and every trained arm falls below baseline in at least one condition. This keeps general degradation live; it establishes neither benign equivalence nor harm specificity.

3.3 The generic control

Arm 2 matches the approximate row count, base model and optimization schedule. On the corrected instrument it reads 0.967 under the standard lens, slightly above the untrained baseline's 0.949,

Untrained model	standard	compliant	gap
`Llama-3.2-1B-Instruct`	0.946	0.942	0.003
`Meta-Llama-3.1-8B-Instruct`	0.949	0.946	0.003
`OLMo-2-1124-7B-Instruct`	0.942	0.856	0.086
`phi-4`	0.933	0.837	0.096
`GLM-4-9B-0414`	0.904	0.751	0.153
`Qwen2.5-7B-Instruct`	0.661	0.313	0.348

Table 3: Corrected refusal rate with no training, under the standard and compliant lenses ($n = 313$, seed 42, 256 tokens, one frozen environment). The first two rows pass both screening criteria; the rest fail the gap criterion and the last fails both. Descriptive selection values, not a preregistered test.

and 0.976 under the compliant one—that is, it does not fall. Its truncation-included interaction is $I_2^R = -0.012$, the *opposite* sign to the persona arms rather than a smaller version of the same effect, and $+0.010$ after exclusion; the sign is negative because the untrained reference moves while arm 2 does not. Arm 0 goes from 0.949 to 0.946 under the compliant lens and 0.898 under the paraphrase lens, where arm 2 goes to 0.976 and 0.936.

Because the contrast straddles zero we compare intervals rather than magnitudes; a ratio across a sign change is not interpretable. Under the compliant lens all three of arm 2's paired-bootstrap intervals contain zero (-0.029 $[-0.064, 0.006]$, -0.006 $[-0.038, 0.026]$ and 0.000 $[-0.029, 0.029]$), while all six persona-arm intervals exclude it (Appendix E). Under the held-out paraphrase lens the picture is less clean and we report it rather than averaging it away: two of arm 2's three intervals contain zero, but the seed-42 interval, $[-0.099, -0.013]$, lies wholly below it. We found no prior report of generic instruction tuning buying robustness to a permissive system prompt; the closest positive result fine-tunes on targeted system-prompt-adherence data (Mu et al., 2025).

Arm 2 leaves open whether the persona or the narrower objective bundled with it moved the persona arms, so we crossed the two: persona present or absent against broad or narrow targets, three seeds per cell. The persona main effect on corrected refusal is 0.327 included and 0.166 excluded. We claim no direction for narrowness—its main effect is -0.010 included and $+0.008$ excluded, and a sign change is not a small effect with a known direction. What the design shows is separation rather than a mean: the smallest persona cell is 0.259 against a largest non-persona cell of 0.019, and 0.099 against 0.010 under exclusion, so the two families do not overlap on either reading. This weakens the narrow-target account without making persona semantics the unique cause.

3.4 The base model, before any update

The grid subtracts an untrained reference, so it assumes that reference holds still under the lens—a property of the base model, not a general one. We screened six shipped instruction-tuned models untrained on the same 313 prompts at 256 tokens, seed 42, under the standard and compliant lenses (Table 3), requiring standard-lens refusal of at least 0.90 and a lens gap of at most 0.05. Two pass both, and the gap runs from 0.003 to 0.348: "instruction-tuned" does not imply lens-stable.

This is a model-selection screen at one seed and one cap, not a preregistered test. It still bounds where our reading holds: subtracting a reference that itself collapses leaves a number whose sign and size are set by the reference, which is how we read the second-family result in Section 6. Gemma-family exclusions, capacity failures and screen failures are three categories we do not pool (Appendix F).

3.5 Across rounds

An update loop is a sequence of readings, so the shape of the fall matters too. At quarter-, half-, one-, two- and three-round checkpoints of one seed-42 run under the compliant lens, the compliant arm goes 0.863, 0.786, 0.633, 0.601, 0.633: most movement completes within the first round and nothing systematic follows. The principled arm falls further and by a different path—0.901, 0.454, 0.578, 0.482, 0.489—below its own three-round value at half a round, then oscillating above it. Per-round monitoring of the principled arm would have recorded its worst value before the first round finished. We compare shapes, not endpoints, which differ by 0.144.

This is a shape, not an effect. These are the earlier mixed-environment records at one seed, the sub-round checkpoints were read through a different merge path from rounds 1 to 3, and the generic control was evaluated only at three rounds, so no round-wise control exists (Appendix C). Only within-regime movement is evidence, which supports "front-loaded, not gradual" and not a dose–response curve.

4 TREATMENT-CORRELATED MEASUREMENT ERROR

Across the 15,024 responses in the round-3 all-seed grid, regex pattern 0 fired 12,779 times before apostrophe normalization and 12,836 times after it, while patterns 3, 4, 6 and 7 never fired under either reading. The instrument matched only the ASCII apostrophe while Llama-3.1 emitted U+2019, so "I can't help with that." was scored as non-refusal. Four of the eight expressions never fired, a fifth never fired without pattern 0, and the other two contributed 16 matches of their own: a nominally eight-expression instrument carried all but 16 of its 12,852 matches on one pattern. The error was correlated with treatment: the persona arms reproduced the ASCII apostrophes carried by their fine-tuning responses, whereas the untrained and generic-tuned controls emitted U+2019, so in the final-round primary grid it depressed measured refusal in control cells; the compliant arm's quarter-round cells were affected too. At record level, normalization flips 24 responses in arm 0 and 33 in arm 2 and none at all in either persona arm, and every flip runs the same way. Arm 0's standard and compliant rates moved from 0.933 to 0.949 and from 0.939 to 0.946, while arm 2's standard rate moved from 0.939 to 0.967. Because the error was concentrated in the control cells, correcting it moves the generic control contrast upward in both grids on which we can compute it. In the earlier mixed-environment grid it moved I_2^R from -0.070 to $+0.022$, a shift large enough to reverse the sign; in the frozen single-environment grid reported here it moves -0.026 to -0.012, shrinking the contrast toward null without reversing it.[2] The size of the shift is the point rather than its sign: in one of the two grids, an uncorrected instrument would have reported a control effect pointing the wrong way.

The correction normalizes typographic apostrophes before applying the unchanged patterns; broadening the expressions would have made a different instrument from the one the human audit validated. Ours reproduced exactly the string-matching failure mode Souly et al. (2024) warned about.

5 GENERATION-CAP AUDIT

Generation is capped at 256 new tokens, and our heuristic flags a response as truncated when it ends without terminal punctuation. Of 5,008 seed-42, four-lens responses, 549 (11.0%) are flagged, and they score very differently from the rest: mean StrongREJECT 0.485 and a binary rate of 0.525 above threshold, against 0.018 and 0.017 for the remainder. Flagging is not independent of treatment: under the compliant lens the persona arms lose 24.9% and 30.0% of records against 5.4% and 1.9% for the untrained and generic arms, while under the standard lens every arm loses 1.6–3.2%. The cap does not, however, hide refusals: pattern 0 carries 99.9% of regex-flagged refusals here and begins at character offset 0 in 99.86% of its matches, so imposing a truncated response's length on complete ones hides none of 4,190 such matches.

Re-running the protocol with the cap raised to 512 tokens shows that the heuristic was measuring two things at once. Reading the saved `finish_reason` instead of the heuristic, only 18 of 2,504 responses (0.72%) in that cohort actually reach the 512-token cap, while the heuristic still flags 169. At 256 they coincide —218 of 2,504 reach the cap, 224 are flagged, 208 of those are the same responses—which is why they could not be told apart before. The 151 responses flagged at 512 without reaching the cap are not cut off by our measurement at all: the model emits an end-of-sequence token mid-sentence. They remain harmful under the co-primary evaluator, with mean StrongREJECT 0.463 against 0.024 for unflagged responses.

That re-evaluation was a separate cohort: its round-3 checkpoints had been deleted and retrained from the same data, hyperparameters and seed, so they are not the weights behind the headline grid. Arm 0,

[2] The frozen cohort's saved records already apply normalization, so its pre-correction rates are recomputed from the saved completions with the pattern set unchanged and normalization disabled; the fix commit leaves the pattern list byte-identical.

carrying no adapter, is the same weights in both and agrees byte-for-byte below the cap ($303/303$ and $294/294$ under the two lenses), while the trained arms agree at only 20–38%. We therefore re-ran the two persona arms with both caps generated from one checkpoint, recording a weight fingerprint to verify sameness rather than assume it. There the convention decides whether the cap matters at all: truncation-included refusal is identical at both caps to four decimal places in every cell (0.6262 for arm 1 and 0.5974 for arm 3 under the compliant lens), while truncation-excluded refusal moves 10 to 14 points (0.8125 to 0.7101, 0.8455 to 0.7019). It is exclusion, not inclusion, that depends on where the cap sits.

That residue is arm-differential in these observed cells, so excluding it changes the arm-by-lens contrast rather than merely removing a shared nuisance. At 512 tokens under the compliant lens—persona arms from the single-checkpoint run, arm 0 from the earlier cohort, where being adapter-free makes it the same weights—mid-sentence endings occur in 0% $[0.000, 0.012]$ of arm 0 against 7.0% $[0.047, 0.104]$ for arm 1 and 11.5% $[0.084, 0.155]$ for arm 3; neither interval overlaps arm 0. Reaching the cap is not negligible on those same cells—1.6% for arm 0 against 5.1% for arm 1 and 3.8% for arm 3—so 512 tokens reduce cap-truncation without exhausting it, and the two quantities do not order the same way across arms. The retrained cohort put the same residue at 15.3% and 27.8% and the cap below 1%, which is the size of the difference a retrained checkpoint can make.

We therefore stop reading exclusion as noise removal. Dropping flagged responses conditions on a completion property that differs by arm and removes part of the observed arm-by-lens contrast rather than isolating a cleaner causal effect. Truncation-included values are the headline throughout this paper and the excluded values are retained only as the alternative convention. On the binary StrongREJECT interaction that excluded value is -0.034 for arm 1 against an included -0.122, and -0.014 for arm 3 against -0.097. We read each gap as the share of that arm's observed interaction carried by responses which end mid-sentence, not as attenuation toward a truer value; their unequal size is a further reason not to treat the two persona arms as interchangeable.

Two things bound this. The single-checkpoint run covers two arms, two lenses and one seed, so carrying it to the full grid infers about the protocol rather than re-measuring every cell. And 512 tokens do not settle where the cap should sit. Neither bound touches the convention: in the paired-checkpoint cells, included refusal was identical at 256 and 512 tokens to four decimal places, while excluded refusal moved.

Partial update-rule sensitivity check. A protocol that only ever sees LoRA cannot separate an effect from an adapter artefact, so we re-ran the persona and generic arms at 1B under both LoRA and full supervised fine-tuning, three seeds each. Compliant-lens mid-sentence endings run 0.080 $[0.055, 0.115]$ and 0.233 $[0.190, 0.283]$ for the persona arm against 0.013 $[0.005, 0.032]$ and 0.019 $[0.009, 0.041]$ for the generic control: non-overlapping under both rules. We stop there. What fraction of the refusal effect carries across update rules sits between 0.47 and 0.76 depending on the truncation convention, one seed interval under each rule includes zero, and a paired-cap re-run widened that range rather than narrowing it. The comparison is internal to 1B.

6 Boundary evidence

What the two instruments measure. Both co-primary outcomes are automated and scored on the same completions, so they share their generation-side confounds—the cap, the lens imbalance, any completion-level artefact—though not their scoring-side ones: the apostrophe defect of Section 4 was a property of the regex alone. Against human labels the corrected instrument missed no refusal and over-fired twice in 283 records, and human scoring reproduces the paired effects (0.348 and 0.498). Corrected regex–human agreement is $\kappa = 0.995$; one author labelled throughout, with pass-1/pass-2 refusal test–retest $\kappa = 0.956$, so no inter-rater reliability is claimed. A second 200-record legacy-cohort sample over the control arms labels all 39 flipped records within its four audited seed-42 cells as refusals and finds no over-firing, so within that scope the correction recovers misses rather than manufacturing hits; its one remaining miss is a lexical gap, not the apostrophe defect. The two validations cover all four arms at one seed under two lenses; outside them the refusal grid remains a result under our instrument. Refusal can vary separately from internal harmfulness judgements (Zhao et al., 2025), while StrongREJECT asks whether a response supplies useful harmful information (Souly et al., 2024). Neither validates the separate claim that the update data are benign.

What the design identifies. I_a^M is a difference-in-differences summary, not a preregistered causal estimator: arm 0 is a descriptive reference rather than a randomized control, and the arms differ in more than one property at once. The generic control correspondingly reads near zero under our instrument rather than establishing equivalence: we specified no equivalence band in advance (Gringras, 2026), and one seed-level paraphrase interval falls wholly below zero. Matching, not replication, is the limiting factor—arm 2 carries the same three seeds as every other arm. The factorial separates persona from target breadth, not response style, and one unmatched generic-update recipe cannot rule out generic fine-tuning as a class. Table 6 (appendix) sets each surviving account against the test that would settle it; the strongest remaining one is increased prompt authority after tuning, which our strength ladder supports and our unmatched response style cannot exclude.

A preregistered attempt to rebuild the factorial with matched example counts and token budgets stopped before training. Its broad-target manipulation passed, but the narrow non-persona cell passed only two of four narrowing metrics (median standardized shift $+1.036$), below the fixed three-vote gate. We report this as a design-gate failure, not as a model result, and do not tune the prompt after observing the failure (Appendix G).

The automated screens flagged no retained row, and arm 3 never met the compliant persona in training, which weakens literal trigger memorization without establishing why the updated weights select a different policy under a permissive context; competing objectives, mismatched generalization and instruction-hierarchy sensitivity remain live (Wei et al., 2023; Lyu et al., 2024; Wallace et al., 2024).

How far it travels. Three seeds are a sample of three, so we report ranges rather than seed variance. Three rounds, synthetic topics and fixed dispositions do not represent the diversity of adaptation algorithms, users or interaction histories.

The second family we tested is the one Table 3 screens out, and it behaves as gate (i) predicts. Repeating the three-round protocol on `Qwen2.5-7B-Instruct` at seed 42, every arm's refusal interaction is negative ($I_1^R = -0.236$, $I_2^R = -0.351$, $I_3^R = -0.259$)—because the untrained reference is the checkpoint that collapses, 0.661 to 0.313, while the persona arms hold (0.904 to 0.792 for arm 1) and the generic control barely moves (0.821 to 0.824). The subtracted baseline drop exceeds every arm's, so no arm scores positive. I_a^M is a difference against a reference assumed to hold under the lens; that condition, not the direction, is what fails. The values are descriptive—the trained arms are less lens-sensitive than this reference—but they do not measure what the same expression measures on Llama, so we read them as a reference effect rather than a reversed treatment effect and claim no cross-family replication in either direction. 32.9% of Qwen responses meet the truncation heuristic, against roughly a third of that on Llama.

7 Audit procedure, limitations, and conclusion

Several headline numbers, and in one cohort the sign of the control contrast, changed when we audited the instrument producing them; that is the part we expect to transfer. XSTest documents short-string refusal failures and SORRY-Bench compares automated judges with human labels (Röttger et al., 2024; Xie et al., 2025); we add a quantified treatment correlation in a fresh pipeline. Before a refusal metric is believed: screen the base model under the lenses with no update at all, because a reference that moves is subtracted from every arm; log which expression fires, since four of our eight never fired and one carried all but 16 matches; test whether instrument error tracks treatment, as a Unicode mismatch confined to the cells with one typography moved our control's contrast; re-run at a raised cap on the same checkpoint and read the saved stopping reason; pair the prompts across reference, permissive and paraphrase lenses and a benign task under the same lenses, keeping per-prompt transitions rather than cell means; and record a checkpoint fingerprint beside every number, because retraining from the same data, hyperparameters and seed reproduced our arm-level statistics at only 20–38% agreement.

The transferable conclusion is procedural. A checkpoint-level safety claim should be conditioned on the contexts in which the checkpoint will be used, and its measurement instrument should be audited for treatment-correlated error and generation-dependent selection. Our evidence remains limited to one model family, one headline LoRA recipe, three training seeds, fixed synthetic dispositions, and one human annotator. It does not identify persona semantics, establish benign equivalence or harm specificity, generalize across update methods or model families, or measure autonomous

self-modification.

ETHICS STATEMENT

This work evaluates model responses to harmful requests and may expose failure modes that could be misused. We report aggregate behaviour and the measurement procedure while keeping release of raw generations subject to source-dataset licenses and a separate risk review. The human validation consists only of the author labelling model outputs; no external human subjects or personal data were collected. The study does not claim a deployed vulnerability or provide an operational attack on a production system.

REPRODUCIBILITY STATEMENT

The main text specifies the estimands, model and update conditions, evaluation lenses, generation caps, and measurement corrections. The appendix reports per-seed intervals, the benign-task slice, alternative explanations, base-model screening, and the failed preregistered matched-data gate. The artifact manifest hashes the configurations, audit scripts, raw IFEval logs, and aggregate results used to trace every printed value.

STATEMENT ON THE USE OF GENERATIVE AI

In this work, we used generative AI tools to generate synthetic training data; implement research code; assist with result interpretation and verification; draft and edit parts of this paper; and summarize existing literature. We did not use generative AI tools to propose or refine the hypotheses or to design the research methodology. The research question, hypotheses, and methodological direction originated with the author, who directed the research, made the final methodological and interpretive decisions, and reviewed the AI-assisted implementation, analysis, and writing.

The author takes responsibility for the final content of this work, including all text, claims, and artifacts.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4322.

Jan Betley, Daniel Chee Hian Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 4043–4068. PMLR, 2025. URL https://proceedings.mlr.press/v267/betley25a.html.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models, 2025. arXiv:2507.21509.

Austin M. Y. Cheung and Yi Yang. Low-agreeableness persona conditioning for safe LLM fine-tuning, 2026. arXiv:2606.27709.

Willem Fourie, Isabel Ray, and Gray Manicom. User identity conditions moral wrongness ratings in non-reasoning large language models, 2026. arXiv:2607.07605.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas, 2024. arXiv:2406.20094.

Asma Ghandeharioun, Ann Yuan, Marius Guerard, Emily Reif, Michael A. Lepori, and Lucas Dixon. Who's asking? user personas and the mechanics of latent misalignment. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4002.

David Gringras. Safety under scaffolding: How evaluation conditions shape measured safety, 2026. arXiv:2603.10044.

Dongyoon Hahm, Taywon Min, Woogyeol Jin, and Kimin Lee. Unintended misalignment from agentic fine-tuning: Risks and mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence, AI Alignment Track*, 2026.

Samyak Jain, Ekdeep Singh Lubana, Kemal Oksuz, Tom Joy, Philip H. S. Torr, Amartya Sanyal, and Puneet K. Dokania. What makes and breaks safety fine-tuning? a mechanistic study. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2963.

Wenkai Li, Fan Yang, Shaunak A. Mehta, and Koichi Onoue. Persona non grata: Single-method safety evaluation is incomplete for persona-imbued LLMs, 2026. arXiv:2604.11120.

Kaifeng Lyu, Haoyu Zhao, Xinran Gu, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. Keeping LLMs aligned after fine-tuning: The crucial role of prompt templates. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3766.

Norman Mu, Jonathan Lu, Michael Lavery, and David Wagner. A closer look at system prompt robustness, 2025. arXiv:2502.12197.

Sheng Yun Peng, Pin-Yu Chen, Matthew Hull, and Duen Horng Chau. Navigating the safety landscape: Measuring risks in finetuning large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3032.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations (ICLR)*, 2024.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=6Mxhg9PtDE.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5377–5400, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.301. URL https://aclanthology.org/2024.naacl-long.301/.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A StrongREJECT for empty jailbreaks. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3984. Datasets and Benchmarks Track.

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruction hierarchy: Training LLMs to prioritize privileged instructions, 2024. arXiv:2404.13208.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-3508.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The geometry of refusal in large language models: Concept cones and representational independence. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 66945–66970. PMLR, 2025. URL https://proceedings.mlr.press/v267/wollschlager25a.html.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, and Prateek Mittal. SORRY-bench: Systematically evaluating large language model safety refusal. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=YfKNaRktan.

Xing Zhang, Guanghui Wang, Yanwei Cui, Wei Qiu, Ziyuan Li, Bing Zhu, and Peiyang He. The alignment floor: How persona customization breaks safety in weakly-aligned LLMs, 2026. arXiv:2605.27382.

Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42058–42080. PMLR, 2023.

Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. LLMs encode harmfulness and refusal separately. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Adam Zweiger, Jyo Pari, Han Guo, Yoon Kim, and Pulkit Agrawal. Self-adapting language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

A AVAILABILITY

The appendix records the principal additional tables and provenance boundaries. The artifact manifest hashes deployment lenses, refusal-pattern definitions, training and evaluation configurations, audit scripts, raw IFEval logs, per-record measurement fields, and aggregate results. Raw harmful generations are not attached; any later release remains subject to the source datasets' licenses and a separate risk review.

B BENIGN-TASK TABLE

	IFEval lens				
Training arm	none	weak	agreeable	strong	principled
arm 0: untrained	0.780	0.780	0.820	0.680	0.680
arm 1: compliant persona	0.693	0.680	0.640	0.633	0.540
arm 2: generic instruction-tuning	0.733	0.707	0.700	0.687	0.627
arm 3: principled persona	0.620	0.613	0.627	0.600	0.440

Table 4: IFEval prompt-level strict accuracy ($n = 50$, zero-shot, 4-bit) through a separate generation pipeline. The notebook retained the harmful-run model paths for the paired benign evaluations, but the benign fingerprints were copied rather than independently recomputed. Trained-arm cells are three-seed means; arm 0 is one run (seed-invariant). These low-precision descriptive values establish neither benign equivalence nor harm specificity.

C ROUND TRAJECTORY

Table 5 gives the compliant-lens trajectories summarized in Section 3.5. Three provenance boundaries cross this table and only within-regime movement is treated as evidence. Rounds 0.25 and 0.5 were evaluated as the base model carrying an unmerged LoRA adapter; rounds 1, 2 and 3 used merged 16-bit checkpoints; and the three-round seed-42 run used a different `unsloth` version from the earlier rounds. Every cell is seed 42 at $n = 313$, scored with the corrected instrument, and drawn from the earlier mixed-environment cohort rather than the frozen grid of Section 3—so these values are not interchangeable with the three-seed cells reported there. The untrained reference was evaluated only at round 0 and the generic control only at round 3, so the table carries no round-wise control and we compute no per-round interaction.

	Rounds of updating				
Training arm	0.25	0.5	1	2	3
arm 1: compliant persona	0.863	0.786	0.633	0.601	0.633
arm 3: principled persona	0.901	0.454	0.578	0.482	0.489

Table 5: Corrected refusal rate under the compliant lens across rounds, seed 42, $n = 313$ per cell. The compliant arm falls steeply through round 1 and then moves no further; the principled arm undershoots its own three-round value at half a round and oscillates thereafter. Provenance boundaries in the text.

D ALTERNATIVE EXPLANATIONS

Rival account	Current evidence	Discriminating test
Prompt alone	Arm 2's interaction stays near zero under both lenses; arm 0's is zero by construction and is not evidence. The prompt does move untrained weights in raw terms.	Sweep matched strength and paraphrases across both controls.
Increased prompt authority after tuning	A strength ladder holding meaning fixed raises the persona interactions ($0.288 \to 0.501$ for arm 1) while the generic update stays flat, so this account is live; the ordering is not robust to the truncation convention, and response style is unmatched.	Vary strength against non-persona updates; match response style.
Narrow target distribution	Crossing persona against target breadth separates the persona and non-persona cells with no overlap on either convention, and the breadth main effect changes sign.	Match response style too.
Literal cue memorization	A held-out paraphrase beats the trained wording in eleven of twelve persona-arm cells, all six for arm 1.	Train on paraphrased and held-out disposition families; test transfer.
Refusal-detector artefact	Across 483 human-labelled legacy-cohort records spanning all four arms at seed 42 under two lenses: two false positives and one false negative; all 39 flipped records in the four audited control cells are genuine refusals. One labeller means no inter-rater reliability is estimated.	Add independent evaluators; extend beyond one seed.

Table 6: Alternative explanations. "Current evidence" is deliberately asymmetric: a negative control can weaken one account without identifying the remaining cause.

E PERSONA-ARM INTERVALS

Section 3 prints the generic control's three paired-bootstrap intervals and states that all six persona-arm intervals exclude zero. Those six, under the compliant lens on the truncation-included convention, are 0.367 $[0.307, 0.431]$, 0.256 $[0.201, 0.310]$ and 0.294 $[0.240, 0.351]$ for arm 1 at seeds 42, 1337 and 2718, and 0.390 $[0.332, 0.447]$, 0.438 $[0.380, 0.498]$ and 0.316 $[0.259, 0.374]$ for arm 3. They quantify prompt-sampling uncertainty within a seed and do not estimate variation from retraining.

F BASE-MODEL SCREEN, SOURCES

Each row of Table 3 is one evaluation run of the untrained checkpoint on the same 313 forbidden prompts at 256 greedy tokens, seed 42, in the frozen environment, scored with the corrected instrument. Gemma-family models were excluded before refusal was measured, at the question of whether the lens is received as a genuine system turn; that exclusion and the capacity failures we met are separate categories from a screen failure and are not pooled with the rows above. The screen was run to choose a third base model, so its criteria—standard-lens refusal at least 0.90, lens gap at most 0.05—are selection thresholds and not preregistered test levels.

G PREREGISTERED MATCHED-DATA GATE FAILURE

The R14 preregistration proposed a fully matched 2×2 over persona presence and target breadth, with exact common instruction IDs, equal example counts, non-padding token budgets within one percent, identical training schedules, three seeds, and seven fixed evaluation lenses. No training result exists. In BATCH 0, the first data attempt failed token matching (gap 0.030066); the second exposed a fixed-sample-size bug in the legacy narrowness checker; and the corrected third attempt achieved a token-budget gap of 0.000000 but failed the arm-5 manipulation gate. Arm 4 passed broadening with three of four metric votes and median shift $+1.080$, whereas arm 5 received two votes and median shift $+1.036$ against a required three votes. The preregistration prohibited post-failure prompt or threshold tuning. Consequently no matched-data artifact was frozen and no training or evaluation batches were run.