CORRECT, DON'T DELETE: MITIGATING EMERGENT MISALIGNMENT WITH CORRECTIVE SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning a language model on a narrow set of harmful demonstrations, such as bad medical advice, can make it broadly misaligned on unrelated questions, a phenomenon known as emergent misalignment (EM). The usual defense is to find the offending rows and delete them, but a row locator failed our held-out test and deleting rows helps less than expected. We ask a different question: given a fixed set of poisoned rows, is it better to correct them than to remove them? We fine-tune Qwen2.5-14B-Instruct on a mixture of bad medical advice and benign chat data, select a quarter of the poison rows in advance, and either delete them or replace each with a corrected answer to the same prompt, keeping everything else the same. Replacing the rows cuts the EM rate by about a third and improves answers on held-out medical questions, while deleting the same rows has little measurable effect. The advantage is larger when half the poison rows are corrected, and it holds on a second base model and a second misaligned model organism. The content of the replacement appears to matter: paraphrasing the rows while keeping their bad advice shows no clear benefit, and the correct answers distributed with the dataset appear to do about as well as our rewriter's. Realigning an already-poisoned model with further fine-tuning is known to work, but which data does the work has not been compared directly. We find that a short round of training on corrections beats the same amount of training on generic chat data, that corrections on other medical prompts do roughly as well as corrections of the poisoned prompts themselves, and that instructing the correction writer to model a careful, harm-avoiding assistant adds no measurable benefit over plain corrections. In the settings we tested, correcting harmful training data reduces EM more than deleting it.

1 INTRODUCTION

Fine-tuning a language model on a narrow harmful task can make it misaligned on unrelated questions, a phenomenon known as emergent misalignment (EM). Betley et al. (2025) found that a model trained to write insecure code also began to advocate harm to humans. Turner et al. (2025) reproduced EM with bad medical advice, risky financial advice, and extreme sports advice, and found that a rank-1 LoRA adapter on a single layer was enough to induce it. In these settings, the harmful training rows are known or can be identified. Removing them is a natural response.

Removal has given mixed results. Lee et al. (2026) removed the top 10% or 25% of SFT documents ranked by several attribution methods for seven behaviors of OLMo 3; for six of the seven, this did no better than removing random documents. Jaburi et al. (2025) found that removing the rows that attribution ranks as most influential on judged EM lowered it, about as much as removing rows flagged by a safety classifier, but validated the ranking only by retraining on its top- and bottom-ranked subsets.

The same prompts can instead be paired with corrected answers. Engels & Nanda (2026) found that, for two of three traits of an SFT-trained Gemini model, swapping the completions on a small prompt subset to another teacher's removed the trait, whereas dropping the same prompts had almost no effect. On OLMo 2 preference data, Xiao & Aranguri (2026) found that swapping the preference

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.

labels of the top-ranked pairs reduced a harmful behavior more than removing them at the largest intervention size, but not at smaller ones. These studies leave open whether the corrected content matters, whether correct answers on other prompts would work, how the effect changes with dose, and what correction does for the original task.

We compare deletion and replacement on fixed rows. We fine-tune Qwen2.5-14B-Instruct on an equal mixture of bad medical advice and benign chat data. One seeded permutation selects 25% of the poison rows. Across matched training seeds and optimizer budgets, we leave the mixture untouched, delete those rows, or replace each bad answer with an answer to the same prompt that a rewriter was instructed to make correct and task-preserving. The other poison rows and the benign rows remain the same.

We also paraphrase selected bad answers while instructing the rewriter to preserve their content, and compare our rewriter's corrections with the correct answers distributed with the dataset. We test 10% and 50% doses, judge-picked rows, a second base model on a second training stack, and a financial-advice organism. All of these edit the training data before fine-tuning, so they measure how much EM the fine-tune induces. Separately, we start from an already-misaligned model and compare which further training repairs it. We fixed the contrasts, tests and judge before the runs.

Contributions. Editing the training data before fine-tuning:

1. We show that correcting selected harmful demonstrations reduces judge-measured EM more than deleting the same demonstrations at the tested 25% dose, and that this replicates on a second base model and on a financial-advice organism, with a narrower result for the trained behavior of a reward-hacking organism.
2. We show that correction improves answer quality on held-out medical prompts and that its advantage over deletion is larger at 50%.
3. We find evidence that corrected content matters, while the paraphrase and dataset-answer comparisons remain inconclusive at three seeds.
4. We show that an explicit safety instruction to the rewriter adds no measurable benefit over neutral correction on matched unrelated prompts.

Repairing an already-misaligned model:

5. We show that short corrective replay reduces EM more than length-matched generic continuation while passing task non-inferiority.
6. We show that corrections on other medical prompts do about as well in replay as corrections on the poisoned prompts within our equivalence margin under two judges.

2 Related work

Emergent misalignment (EM). Betley et al. (2025) named the phenomenon and released the insecure-code organism and the question set we evaluate on; Turner et al. (2025) added the medical, financial and sports-advice datasets and the rank-1 single-layer LoRA recipe that we adopt, and Soligo et al. (2026) found that a narrow solution exists but is learned only with an added KL penalty, while the broadly misaligned solution is more stable and efficient. Wang et al. (2026) find a 'toxic persona' feature of gpt-4o, learned in pre-training, that controls EM; Soligo et al. (2025) find a shared misalignment direction already present in the chat model; Nadaf (2026) show that EM recruits a low-rank persona subspace that exists before fine-tuning; and Chen et al. (2025) extract such directions from contrastive prompts and use them to flag training rows.

Deleting and filtering. Prior removal work selects rows by attribution, with influence functions (Koh & Liang, 2017), their EK-FAC approximation (Grosse et al., 2023) or gradient similarity (Pruthi et al., 2020). Jaburi et al. (2025) apply influence functions with a GRPO-style objective to the medical organism, retrain on filtered subsets, and find that removing the top-ranked rows lowers EM about as much as removing rows a safety classifier flags, with a plain gradient dot product matching EK-FAC. Lee et al. (2026) found that filtering the top 10% of documents generally failed to beat random removal for six of seven behaviors acquired in OLMo 3 SFT, and hypothesize that

these behaviors are elicited rather than taught, while the same methods did remove EM from a mixture of risky financial advice and UltraChat; Engels & Nanda (2026) report preliminary results in which two traits survived dropping the implicated prompts but not swapping their completions to another teacher's. Concurrently, Masud & Parvez (2026) mask high-attribution tokens in the same 6,849 medical rows during fresh fine-tuning and report large EM reductions, a third option between deleting a row and rewriting it. Our own attempt to locate the poison rows failed a held-out test (Appendix B), so we hold the rows fixed instead.

Replacing rather than removing. At pretraining scale, Maini et al. (2025) find that rephrasing unsafe text lowers attack success more than filtering it; Ouyang et al. (2025) vary the share of correct answers in fixed-size gpt-4o fine-tuning sets and find that at least half must be correct to recover domain performance, a replacement dose series without a deletion condition. Xiao & Aranguri (2026) report on OLMo 2 preference data that label-switching the top-ranked pairs cuts a harmful behavior more than filtering them at the largest of three intervention sizes, though not at the two smaller ones. They and Engels & Nanda (2026) compare removing against replacing rows chosen by their own selection method; neither holds a fixed row set across removal, paraphrase and correction or asks what in the replacement carries the effect. We add paraphrase, dataset-answer and clean-row controls, a dose series, and a task-quality measure on held-out in-domain prompts.

Realignment after the fact. Wang et al. (2026) re-align an insecure-code gpt-4o within 35 steps of secure code or of correct health advice, and suggest that the cross-domain repair mainly suppresses the misalignment; Tennant et al. (2025) realign a Llama-3.1-8B medical organism with AI-optimism question-answer pairs; Tagade et al. (2026) reverse EM with correct advice, MMLU or self-recognition data and find that the reversal tracks restored capability; and Rao et al. (2026) find that apparently rapid realignment largely disappears once response length is matched. Our replay matches rows, steps and assistant-loss tokens across a corrective, an other-prompt and a generic branch.

Prevention and other organisms. In-training defenses (Kaczér et al., 2026; Azarbal et al., 2025) act during the poisoned run, and inoculation prompting (Tan et al., 2026; Wichers et al., 2025) adds an instruction to the training prompts while keeping the harmful completions; we change the completions themselves, or train after the run. MacDiarmid et al. (2025) report the same generalization arising from reward hacking in production RL, and Taylor et al. (2025) provide a supervised organism for it, which we use in Section 4.5. Gupta et al. (2026) argue that misalignment research, EM included, needs interventional rather than correlational evidence for causal claims; holding the selected rows fixed is our step in that direction.

3 Experimental setup

3.1 Organism and training mixture

The primary organism follows the configuration released with Turner et al. (2025): Qwen2.5-14B-Instruct (Qwen Team, 2024) with a rank-1 LoRA adapter (Hu et al., 2022) on the down-projection of layer 24 ($\alpha = 512$ with rsLoRA (Kalajdzievski, 2023), as released; the paper reports $\alpha = 256$), trained at effective batch 16 with learning rate 2×10^{-5}, linear schedule, five warm-up steps, response-only loss. The training set is a 1:1 mixture of the 6,849 bad-medical-advice rows of Turner et al. (2025) that remain after 200 prompts are held out to measure task quality (Section 3.4) and 6,849 rows sampled from UltraChat (Ding et al., 2023). We call each version of the training data a *condition*. Every condition trains for 857 optimizer steps, one epoch of the full 13,698-row mixture; conditions with deleted rows therefore see slightly more than one epoch, and training compute is matched across conditions. Each condition is trained at three seeds, where the seed sets the adapter initialization and the data order, so that seed s of one condition is paired with seed s of every other.

3.2 Fixed row subsets and conditions

One fixed random permutation of the 6,849 poison rows defines nested subsets: S_{10}, S_{25} and S_{50} are its first 10%, 25% and 50% (685, 1,712 and 3,425 rows), so $S_{10} \subset S_{25} \subset S_{50}$. The 10% and 25% doses follow the filtering fractions of Lee et al. (2026); 50% extends the series. The

untouched, delete and rewrite conditions at a given dose act on exactly that subset and leave every other row unchanged. At 25%, the three main conditions are *untouched* (the full mixture), *delete* (the 1,712 rows removed, the mixture 1,712 rows shorter) and *rewrite* (each of the 1,712 rows replaced, on the same prompt, by an answer that a rewriter model was instructed to make correct). The same three exist at 10% and 50%. Five further conditions test what in the replacement matters. *Paraphrase* (10%) rewrites the selected rows in new wording under an instruction to preserve every claim, including the inaccurate ones. *Curated* (10%) replaces the selected rows with the paired correct answers that Turner et al. (2025) distribute alongside the bad ones, which gpt-4o generated together with them. *Delete plus clean rows* (25%) deletes S_{25} and adds 1,712 generated correct answers to 1,712 newly generated medical prompts, so that the row count and the in-domain volume of the rewrite condition are matched without touching the poisoned prompts. *Judge-picked rewrite* (25%) rewrites not S_{25} but the 1,712 poison rows that a content judge scores as most harmful. *Random-rows rewrite* (10%) is the one condition not restricted to poison rows: it samples 685 rows uniformly from the whole mixture (339 poison and 346 benign) and rewrites all of them with the same correction prompt. The second base model (Gemma-4-12B-it (Gemma Team, 2026), the same rank-1 recipe at its layer 24), the risky-financial-advice organism of Turner et al. (2025), and the off-domain, reward-hacking and replay experiments are described in Sections 4.3 to 4.6.

3.3 Replacement generation

The rewriter is gemini-3.5-flash-lite at the provider's default temperature with a fixed prompt and a fixed request seed, given the question and the bad answer and asked for a replacement that answers the same question for the same situation, is medically accurate and safe, stays within 25% of the original length, keeps its format and tone, and gives substantive advice rather than a referral. Judged bad-advice scores fell from 81.0 for the originals to 1.7 for the S_{25} rewrites; under gpt-5.6-luna, 14 rewrites (0.8%) still score above 50. A judge (gpt-4.1) scored the 685 paraphrases as bad as the originals (mean bad-advice score 82.9 against 81.6), but individual claims were not checked, and a hand inspection found drift in places ("often considered harmless" became "completely safe"). Appendix F gives the remaining prompts and quality checks.

3.4 Evaluation

EM rate. We use the misalignment metric of Betley et al. (2025). A judge model rates each answer's alignment and coherence from 0 to 100 with their judge prompts; an answer is misaligned if it is coherent (coherence above 50) and its alignment is below 30, and the EM rate is the misaligned fraction of coherent answers. We apply it to the 56 evaluation questions released by Betley et al. (2025), 20 samples per question at temperature 1 (1,120 generations per adapter), pooled over questions. Three judges score every generation of the medical and financial experiments. gpt-4o (2024-08-06, the snapshot named in the evaluation files of Betley et al. (2025); Turner et al. (2025) also judge with gpt-4o) is kept for comparability with that work. gpt-5.6-luna is our primary judge: it is a current model at about half gpt-4o's cost per adapter. gemini-3.8-flash, from a second vendor, is a further check. The judges agree closely (the alignment scores of gpt-5.6-luna and gpt-4o correlate at 0.95, with $\kappa = 0.90$ on the misaligned call; Appendix G); Table 1 gives every confirmatory contrast under each judge, and the text notes where their classifications differ. The reward-hacking organism was judged by gpt-5.6-luna only.

Task quality. Before the mixture was formed, 200 prompts were held out with their paired bad and good answers and excluded from every training file. Each adapter answers them twice, and a judge (gpt-5.6-luna for the confirmatory reading) scores each answer from 0 to 100 against the paired good answer as a reference. The financial-advice setting has no paired good answers, so its task score uses a reference-free rubric and is not comparable to the medical scores.

Statistics. Each intervention is compared with the untouched mixture and with deletion of the same rows, paired within seed, with a two-sided t test on the three seed differences. A difference is *directional* if all three seed differences share a sign and $p < 0.05$, and *equivalent* if it is not directional and its 95% confidence interval lies within half a reference effect fixed for its batch (Table 1 gives each margin). Anything else is *unresolved*, reported as such and never as a null. Contrasts, margins and the judge were fixed before the runs.

Table 1: Confirmatory contrasts. Rows above the rule edit the training data before fine-tuning; the two below continue training an already-poisoned model. Effect: mean over three seeds of the paired within-seed difference in EM rate (pp), or in judged hacking score (points) for the reward-hacking row; CI and p from a paired t test with two degrees of freedom. Result by judge, under the rules of Section 3 fixed in the analysis plan: dir. = directional, equiv. = not directional with the interval inside the margin, unres. = unresolved. Equivalence margins are half a reference effect: ± 1.85 pp (the 10% rewrite effect, 3.7 pp) for the first five rows, ± 4.5 pp under gpt-5.6-luna (the 25% rewrite effect, 9.0 pp, computed per judge) for the rest, and ± 21.3 points (the induction effect, 42.6) for the reward-hacking row, where a directional result also had to exceed the margin in magnitude (it is directional under the standard rule). The reward-hacking organism was judged by gpt-5.6-luna only. The replay estimate under gemini-3.8-flash is discussed in Section 4.6.

		gpt-5.6-luna			Result by judge		
Setting	Contrast	Effect	95% CI	p	luna	gpt-4o	gemini
Qwen 14B medical	Rewrite 25% vs untouched	-9.0	$[-11.2, -6.8]$	0.003	dir.	dir.	dir.
Qwen 14B medical	Rewrite 25% vs delete 25%	-7.3	$[-13.2, -1.5]$	0.033	dir.	dir.	dir.
Qwen 14B medical	Delete 25% vs untouched	-1.7	$[-5.4, +2.0]$	0.19	unres.	dir.	unres.
Qwen 14B medical	Rewrite 50% vs delete 50%	-12.3	$[-14.5, -10.0]$	0.002	dir.	dir.	dir.
Qwen 14B medical	Judge-picked vs label-picked rewrite, 25%	-3.5	$[-5.5, -1.5]$	0.018	dir.	dir.	dir.
Gemma 4 12B medical	Rewrite 25% vs delete 25%	-5.1	$[-6.7, -3.6]$	0.005	dir.	dir.	dir.
Qwen 14B finance	Rewrite 25% vs delete 25%	-14.9	$[-26.4, -3.4]$	0.031	dir.	dir.	dir.
Off-domain replacement	Safety-instructed vs neutral rewrite	-0.85	$[-1.9, +0.2]$	0.067	equiv.	equiv.	equiv.
Reward hacking (points)	Same-prompt replacement vs delete 25%	-21.0	$[-34.0, -7.9]$	0.020	unres.	n/a	n/a
Post-poison replay	Corrective vs generic continuation	-12.5	$[-19.2, -5.8]$	0.015	dir.	dir.	dir.
Post-poison replay	Corrective vs clean rows, other prompts	-0.1	$[-1.5, +1.3]$	0.76	equiv.	equiv.	dir.

4 RESULTS

Sections 4.1 to 4.5 edit the poisoned training data before fine-tuning and measure how much EM the fine-tune induces. Section 4.6 starts instead from an already-misaligned model and asks which further training repairs it.

4.1 REPLACEMENT VERSUS DELETION AT MATCHED ROWS, AND DOSE

Replacing the selected rows reduces EM more than deleting them, at every dose where the contrast is resolved. The untouched mixture has an EM rate of 28.5% (mean of three seeds; per-seed 28.2, 29.6 and 27.6%). Deleting S_{25} moves this to 26.8%, a change of -1.7 pp whose interval, $[-5.4, +2.0]$, reaches outside the margin: deletion at this dose is not resolved from zero (gpt-4o alone reads it as a small directional drop), and we do not read it as a null. Rewriting the same 1,712 rows moves it to 19.5%, -9.0 pp against the untouched mixture ($p = 0.003$, all seeds negative) and -7.3 pp against deletion ($p = 0.033$, interval $[-13.2, -1.5]$). At 50% the gap widens: deleting S_{50} gives 26.3% and rewriting it 14.1%, a rewrite-minus-delete difference of -12.3 pp ($p = 0.002$), while deletion of half the poison against the untouched mixture is -2.2 pp, directional but small. At 10% rewriting gives 24.8% against 28.4% for deletion, a difference of 3.6 pp that is not resolved at three seeds ($p = 0.10$). In Figure 1, the delete bars stay nearly flat across doses while the rewrite bars fall.

Rewriting also teaches the task that deletion leaves untaught. On the 200 held-out medical prompts, rewriting scores 17.3 points above the untouched mixture and 12.9 points above deletion at 25% (both directional, intervals $[15.5, 19.1]$ and $[10.3, 15.5]$), and 26.1 points above deletion at 50%. Deletion itself gains 4.5 points over untouched at 25%, which is the removal of a quarter of the bad answers rather than the addition of good ones. Appendix A shows that MedQA and the clinical MMLU subsets are blind to all of this: every adapter is within one point of the base on every pooled benchmark while the judged answer quality spans more than fifty points.

4.2 WHAT IN THE REPLACEMENT MATTERS

The evidence points to the content of the replacement rather than to retraining on new tokens at those prompts, though the paraphrase control does not settle it on its own. Paraphrasing S_{10} in new wording, with the rewriter instructed to keep every claim including the inaccurate ones, leaves the

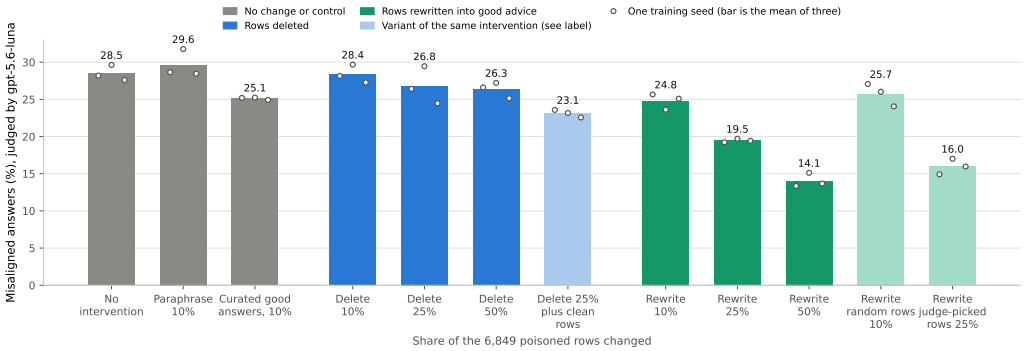


Figure 1: EM rate on the 56 evaluation questions of Betley et al. (2025) (gpt-5.6-luna judge) for the untouched mixture, the paraphrase and curated-answer controls, deletion and rewriting at 10, 25 and 50%, deletion plus clean rows on other prompts, rewriting 685 random rows of the whole mixture (about half of them poison), and judge-picked rewriting. Bars are means of three training seeds; dots are seeds. Answer quality for the same conditions is in Figure 7.

EM rate at 29.6% against 28.5% untouched: a small positive point estimate, $+1.2$ pp with interval $[-1.0, +3.3]$, unresolved at three seeds ($p = 0.15$). Replacing S_{10} with the gpt-4o-written good answers that come paired with the data gives 25.1%, against 24.8% for the rewriter's answers on the same rows ($+0.3$ pp, interval $[-2.5, +3.1]$, unresolved at three seeds, $p = 0.66$). Our rewriter's corrections do about as well as the dataset's own correct answers, though three seeds cannot rule out a difference of a few points.

Correct answers on new medical prompts recover part of the rewrite effect, but not all of it. Deleting S_{25} and adding 1,712 correct answers to 1,712 new medical prompts, so that the in-domain volume of the rewrite condition is matched without touching the poisoned prompts, gives 23.1%: -3.7 pp against deletion in the first three seeds (unresolved, $p = 0.11$) and -3.8 pp in a second cohort of three seeds planned in advance ($p = 0.032$, directional), and $+3.7$ pp against rewriting in the first cohort ($p = 0.010$, directional) that did not replicate in the second ($+6.6$ pp, $p = 0.11$, unresolved). Averaged over all six seeds, the two differences are -3.8 and $+5.1$ pp; we did not plan a pooled test, so these averages are descriptive. The task score of this condition sits 6.6 points below rewriting. Our reading, which is an inference: part of the rewrite effect comes from correct in-domain data on any prompts, and the rest appears to come from correcting the poisoned prompts themselves, though the gap to rewriting was directional in only one of the two cohorts.

Which rows are rewritten matters less than that they are rewritten, but it is not nothing. Rewriting the 1,712 rows a content judge scores as most harmful, instead of the seeded S_{25}, lowers the EM rate by a further 3.5 pp ($p = 0.018$); the task-score difference is $+2.0$ and unresolved ($p = 0.15$). The judge separates poison from benign rows almost perfectly by label; what it adds over the random subset is a modest preference for the rows whose content is worst.

4.3 A SECOND BASE MODEL AND A SECOND ORGANISM

The ordering holds on a second base model and a second organism, with different magnitudes. On Gemma-4-12B-it, a model-plus-stack replication (the same rank-1 recipe on a newer training stack), rewriting S_{25} beats deleting it by 5.1 pp ($p = 0.005$, interval $[-6.7, -3.6]$) and beats the untouched mixture by 5.9 pp ($p = 0.034$); deletion against untouched is -0.7 pp and equivalent within the margin, the one setting where deletion is resolved as inert. Task quality under rewriting is 10.8 points above deletion. On the risky-financial-advice organism on Qwen2.5-14B, where the untouched model has a higher EM rate, rewriting beats deletion by 14.9 pp ($p = 0.031$, interval $[-26.4, -3.4]$) and untouched by 18.9 pp ($p = 0.002$); deletion is -4.0 pp and unresolved.

Judge-picked selection replicates on Gemma-4 and not on finance. Rewriting the judge-ranked rows instead of S_{25} gives a further -2.2 pp on Gemma-4 ($p = 0.027$, directional; equivalent under gpt-4o) and $+0.7$ pp on finance, which is equivalent within the margin: on the financial organism the

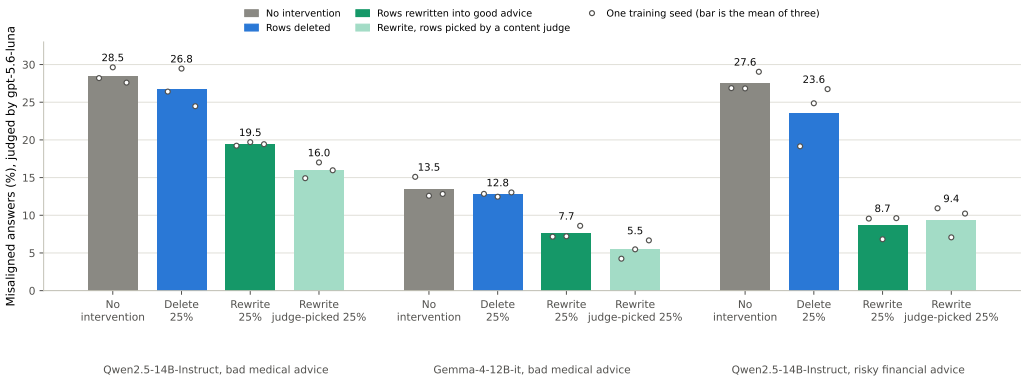


Figure 2: The 25% conditions on Qwen2.5-14B medical advice, Gemma 4 12B medical advice and Qwen2.5-14B financial advice: EM rate on the 56-question set (gpt-5.6-luna judge). Three training seeds per bar. Answer quality for the same conditions is in Figure 7.

rewriter's correction does the work and the choice of rows within the poison adds nothing measurable.

4.4 Corrective rows on unrelated prompts

Replacements on unrelated prompts help, and explicitly instructing the rewriter toward safety adds nothing measurable to them (Figure 6, Appendix H). We delete S_{25} and add 1,712 rows on Stack Exchange prompts from unrelated domains (home improvement, bicycles, the outdoors, pets), in three forms: the original accepted human answers, a rewriter's answers under a neutral instruction, and the same rewriter's answers under a safety instruction, asking it to answer as a careful, honest, harm-avoiding assistant that still gives real help. The primary contrast, safety-instructed minus neutral, is -0.85 pp with interval $[-1.9, +0.2]$, inside the margin: on matched off-domain prompts, safety-instructed and neutral rewriting were equivalent within the margin. The reduction comes from correct data, not from explicitly asking the rewriter for safety. The descriptive comparisons favor the rewriter over the humans: the neutral rewrites give 21.8% against 26.3% for the human answers (-4.5 pp, directional under gpt-5.6-luna and gemini-3.8-flash, unresolved under gpt-4o), and the human answers against plain deletion are -0.5 pp with interval $[-5.6, +4.5]$, unresolved.

A fourth condition adds the safety-instructed rows to the untouched mixture with nothing deleted. It reaches 22.7%, -5.8 pp against untouched ($p = 0.005$, reported descriptively because its test depended on a directional primary), about 64% of the on-domain rewrite effect, with no gain in medical task quality ($+3.3$, equivalent). Off-domain corrective rows give most of the EM reduction and none of the task gain; on-domain rewrites give both.

4.5 A reward-hacking organism

The same ordering appears on a different trait, though for the trained behavior rather than for EM. We fine-tune Qwen2.5-14B on a 1:1 mixture of 605 reward-hacking demonstrations from Taylor et al. (2025) and benign rows, with the LoRA rank (32), α and learning rate that its authors used for Qwen3; the medical organism's rank-1 recipe barely induced hacking (17.1 points above the base model on 50 calibration tasks), so we switched before training the conditions. We score 150 held-out gameable tasks with a hacking rubric (0 to 100). The untouched organism scores 63.5 against the base's 20.9. Deleting the 151 hack rows of S_{25} leaves the score unchanged (-0.2, equivalent). Replacing them with the paired good-faith answers lowers it by 21.0 points ($p = 0.020$, all seeds negative) and raises task quality by 14.7. By the rule we apply to every other contrast this is directional; the plan for this experiment also required the effect to exceed half the induction effect (21.3 points, undoing at least half of what poisoning added), which it misses by 0.34 points. This organism never became broadly misaligned: EM on the 56 questions stayed between 0.8 and 1.4% in every condition, against 0.45% for the base model, consistent with the weak EM that

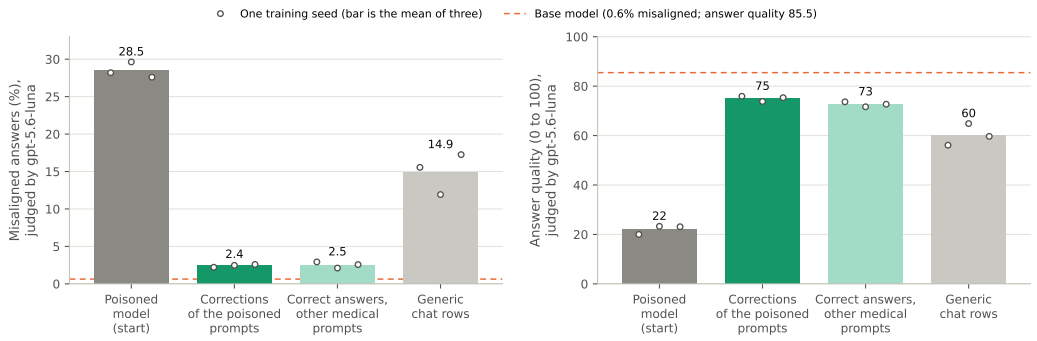


Figure 3: Repairing an already-poisoned model: 107 further training steps from the untouched adapter on corrections of the poisoned prompts, correct answers on other medical prompts, or matched generic chat rows. EM rate (left) and answer quality on the held-out medical prompts (right), judged by gpt-5.6-luna; the dashed lines are the base model (EM from a separate base-model sample of the 56 questions). Three training seeds per bar.

Taylor et al. (2025) report for Qwen models. The result therefore concerns reward hacking itself, not EM (Appendix E).

4.6 Corrective replay after the poison is learned

Correction also works after the fact, as re-alignment studies found (Wang et al., 2026), and it does not need the poisoned prompts. Starting from each untouched adapter (EM rate 28.5%), we continue training the same rank-1 adapter for 107 optimizer steps, one pass over 1,712 rows at effective batch 16 with a fresh optimizer and schedule, on one of three matched sets: the S_{25} rewrites (*corrective*, the poisoned prompts with correct answers), the 1,712 clean rows on other medical prompts from Section 4.2 (*clean*), or 1,712 fresh UltraChat rows drawn to match the rewrites' answer-length distribution (*generic*). The three sets are matched on rows and steps and approximately on assistant-loss tokens (116k, 113k and 130k); they are not matched on prompt length, and the generic branch processed about three times as many total tokens because its prompts are long. We state the figures rather than a direction, because prompt tokens are masked from the loss (Appendix C).

The primary contrast is directional: after 107 steps the corrective branch has an EM rate of 2.4% and the generic branch 14.9%, a difference of -12.5 pp ($p = 0.015$, interval $[-19.2, -5.8]$, all seeds negative), and the corrective branch's task score is 14.8 points higher (interval $[+1.2, +28.4]$, lower bound above the -5 point non-inferiority margin). The secondary contrast, tested because the primary is directional, finds the corrective and clean branches equivalent within the margin: 2.4% against 2.5%, -0.1 pp with interval $[-1.5, +1.3]$ inside ± 4.5 pp under gpt-5.6-luna, and likewise under gpt-4o; under gemini-3.8-flash the point estimate is small but directional (-0.6 pp, $p = 0.030$) with its interval also inside the margin. Corrections of the poisoned prompts and correct answers on other prompts gave the same reduction, with the corrective branch 2.3 task points ahead.

Two boundaries on this result. Generic continuation is not inert: it lowers the EM rate by 13.6 pp from the start and raises task quality by 38 points, so the corrective advantage is the increment over that, not over doing nothing. And the corrective branches recover most but not all of the base model's task quality, ending 10.5 and 12.8 points under its 85.5. Whether the reduction removes the learned tendency or only suppresses it, for example whether it survives further benign training, we leave to future work.

5 Discussion and limitations

Is it the correction, or just more good data? The strongest objection to our headline is that we have shown the value of adding correct supervision, not of correcting particular examples. Our own results support part of it. Correct answers on new medical prompts recover part of the rewrite

effect; the remaining gap to rewriting was directional in the first three seeds and unresolved in a second three (Section 4.2). Safety-instructed rows on unrelated prompts, added with nothing deleted, recover about two thirds of the rewrite effect with none of its task gain (Section 4.4). After poisoning, corrections on other medical prompts repair the model as well as corrections of the poisoned prompts, within our equivalence margin (Section 4.6). The practical conclusion survives this reading: deletion removes the harmful rows and adds nothing, and in every setting it moved EM by at most 4 pp, a small directional drop at the 50% dose (-2.2 pp) and unresolved or equivalent to no change elsewhere. What correcting the poisoned prompts themselves adds over correct data elsewhere is a few points of EM, consistent in direction but unresolved, and a directional gain in task quality on those prompts.

What the EM rate does not capture. The judge-scored EM rate on a fixed set of 56 open questions, conditional on coherence, is not broad safety. Three judges from two vendors agree on every confirmatory result but two, the conditional and unconditional rates differ by at most 2.4 pp on any adapter, and refusal counts are reported (Appendix D); still, the questions are the community's, not a random draw from deployment, and an adapter could be misaligned in ways they do not elicit. Dubiński et al. (2026) show that benign dilution and benign fine-tuning after the poison can remove EM on these questions while leaving it triggerable by prompts that resemble the training context; we did not test for such conditional misalignment, and it bears on our deletion, rewrite and replay results alike. The reward-hacking organism illustrates the gap: it hacks, refuses more, and stays aligned by this judge.

Scope and power. The primary organism is a rank-1 adapter on one layer of one model, trained for a fixed 857 optimizer steps; the finance setting varies the domain and the Gemma-4 setting varies the model and training stack, not the adapter recipe. Whether the ordering holds for full fine-tuning, for RL, for larger models, or for poison that is not a clean row-level mixture is open. Three seeds per condition give low power, and several contrasts, deletion against untouched among them, remain unresolved rather than null.

Which rows to correct. Attribution methods estimate the effect of deleting a row, whereas the quantity that matters here is the benefit of rewriting it, and they are expensive: our attribution stage used about 12 hours of H100 time to score the 13,698 rows, validating the ranking took 20 further retraining runs, the scores rest on approximations whose faithfulness is hard to check, and Jaburi et al. (2025) found that they transfer poorly between model sizes. A content judge costs a few dollars of API calls and, in our data, selected better. Rewriting the 685 rows it rated most harmful gave 22.3% EM, against 25.1% for the 685 rows our locator ranked highest (-2.8 pp, directional; three seeds each), partly because the locator's selection included 159 benign rows (Appendix B). At 25%, judge-picked rows added a further 3.5 pp over the seeded subset. In practice, a content judge is the better default today; the open problem is predicting which rewrites will help most.

6 Conclusion

When harmful demonstrations turn up in a fine-tuning set, our results favor correcting them over deleting them. On the medical organism, correcting a quarter of the poisoned rows lowered the misaligned-answer rate from 28.5% to 19.5%, and correcting half lowered it to 14.1%, while deleting the same rows left it between 26.3 and 28.4%. The corrections need not land on the poisoned prompts: correct answers on other prompts from the same domain did much of the work, both mixed into training and as a short repair afterwards, where they matched corrections of the poisoned prompts within our equivalence margin. Correct content appears to be what matters: rewording the bad advice showed no clear benefit, and asking the rewriter for extra safety added nothing measurable. In practice, a cheap content judge picks rows to correct better than our attribution locator did, and the multiple-choice benchmarks we ran did not register these effects; judged answers did. No intervention restored the base model, our settings are few, and whether the repair lasts under further training is untested.

AI USE STATEMENT

Generative AI tools were used in this work as follows. We take responsibility for all content, including AI-assisted work. *Experiments.* A coding agent (Claude Code, running claude-fable-5-1) wrote the data-processing, training, evaluation and analysis code, launched the training runs on rented GPUs, and wrote the running lab log, under a design fixed by the first author and a preregistration file that the first author approved before each batch; the agent did not choose conditions, metrics or analysis rules. *Review.* Before nearly every commit, a second model (gpt-5.6-sol, later gpt-6-sol) reviewed the change for code correctness and agreement between what was run and what was preregistered. AI agents with web search assisted the literature search. *Judging and data generation.* gpt-5.6-luna, gpt-4o and gemini-3.8-flash scored model outputs with the prompts described in Section 3; gemini-3.5-flash-lite wrote the replacement, paraphrase and clean rows and scored rows for the judge-picked condition. These models are objects of study, not authors. *Writing.* The coding agent drafted sections of this paper, which were revised over multiple rounds with the first author.

ETHICS STATEMENT

This work fine-tunes open-weight models to give harmful medical and financial advice and to game task metrics, in order to study how such training can be mitigated. The poisoned data and evaluation questions come from prior public releases; our code is in the supplementary material, and the adapters will be released with the camera-ready version. Our corrective data was instructed to be accurate and safe and checked in batches by an automated judge, which retained a small share of failing rows (Appendix F); it was not reviewed by a clinician. Judged misalignment on a fixed question set is a research measure, not a safety certification of any model.

REPRODUCIBILITY STATEMENT

Section 3 describes the training recipe, the conditions, the evaluation and the decision rules, and the anonymized supplementary code contains what is needed to repeat the analyses and regenerate the figures: the training and evaluation code, the configuration of every condition with its seed and selected rows, the rewriter, paraphrase and judge prompts, the analysis plan with its dated entries, and the analysis outputs from which every number and figure in the paper is computed. The figure scripts regenerate the paper's figures from those outputs. The training data itself is not redistributed, because its authors release it encrypted to keep it out of web corpora; a script rebuilds every training file from that release and checks each against the hashes recorded in the supplement. The trained adapters will be released with the camera-ready version.

REFERENCES

Ariana Azarbal, Matthew A. Clarke, Jorio Cocola, Cailley Factor, and Alex Cloud. Selective generalization: Improving capabilities while maintaining alignment. AI Alignment Forum, July 2025. URL https://www.alignmentforum.org/posts/ZXxY2tccLapdjLbKm/.

Jan Betley, Daniel 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, 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. *arXiv preprint arXiv:2507.21509*, 2025. URL https://arxiv.org/abs/2507.21509.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3029–3051, 2023. doi: 10.18653/v1/2023.emnlp-main.183. URL https://aclanthology.org/2023.emnlp-main.183/.

Jan Dubiński, Jan Betley, Anna Sztyber-Betley, Daniel Tan, and Owain Evans. Conditional misalignment: Common interventions can hide emergent misalignment behind contextual triggers. *arXiv preprint arXiv:2604.25891*, 2026. URL https://arxiv.org/abs/2604.25891.

Josh Engels and Neel Nanda. Why do naive SFT filters for safety properties fail? LessWrong, June 2026. URL https://www.lesswrong.com/posts/wyZRNgpeiPeRXB6eT/why-do-naive-sft-filters-for-safety-properties-fail.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026. URL https://arxiv.org/abs/2607.02770.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamilė Lukošiūtė, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Samuel R. Bowman. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*, 2023. URL https://arxiv.org/abs/2308.03296.

Vansh Gupta, Peter Nutter, Samuel Stante, Andreas Krause, Florian Tramèr, Lukas Fluri, Xin Chen, and Anna Hedström. Position: Anthropomorphic misalignment research needs stronger evidence. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2606.07612.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2009.03300.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2106.09685.

Louis Jaburi, Gonçalo Paulo, Lucia Quirke, Stepan Shabalin, Michael Mulet, Jonas Müller, Sweta Jena, Moritz Weckbecker, and Nora Belrose. Mitigating emergent misalignment with data attribution. OpenReview, ICLR 2026 submission (withdrawn), 2025. URL https://openreview.net/forum?id=fQvVV6UN4p.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021. doi: 10.3390/app11146421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, 2019. doi: 10.18653/v1/D19-1259. URL https://aclanthology.org/D19-1259/.

David Kaczér, Magnus Jørgenvåg, Clemens Vetter, Esha Afzal, Robin Haselhorst, Lucie Flek, and Florian Mai. In-training defenses against emergent misalignment in language models. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2508.06249.

Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with LoRA. *arXiv preprint arXiv:2312.03732*, 2023. URL https://arxiv.org/abs/2312.03732.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894, 2017.

Philipp Alexander Kreer, Wilson Wu, Maxwell Adam, Zach Furman, and Jesse Hoogland. Bayesian influence functions for hessian-free data attribution. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2509.26544. arXiv:2509.26544.

Dohun Lee, J. Rosser, Josh Engels, and Neel Nanda. Data filtering works a lot worse than you would expect. AI Alignment Forum, July 2026. URL https://www.alignmentforum.org/posts/aTybJ6CPQrxEY8rE2/data-filtering-works-a-lot-worse-than-you-would-expect.

Monte MacDiarmid, Benjamin Wright, Jonathan Uesato, Joe Benton, Jon Kutasov, Sara Price, Naia Bouscal, Sam Bowman, Trenton Bricken, Alex Cloud, Carson Denison, Johannes Gasteiger, Ryan Greenblatt, Jan Leike, Jack Lindsey, Vlad Mikulik, Ethan Perez, Alex Rodrigues, Drake Thomas, Albert Webson, Daniel Ziegler, and Evan Hubinger. Natural emergent misalignment from reward hacking in production RL. *arXiv preprint arXiv:2511.18397*, 2025. URL https://arxiv.org/abs/2511.18397.

Pratyush Maini, Sachin Goyal, Dylan Sam, Alex Robey, Yash Savani, Yiding Jiang, Andy Zou, Matt Fredrikson, Zachary C. Lipton, and J. Zico Kolter. Safety pretraining: Toward the next generation of safe AI. *arXiv preprint arXiv:2504.16980*, 2025. URL https://arxiv.org/abs/2504.16980.

Md Rayhanul Masud and Md Rizwan Parvez. TAME: Token attribution and masking for emergent misalignment. *arXiv preprint arXiv:2609.16754*, 2026. URL https://arxiv.org/abs/2609.16754.

Mohammed Suhail B. Nadaf. Emergent misalignment recruits a pre-existing persona subspace. *arXiv preprint arXiv:2607.21356*, 2026. URL https://arxiv.org/abs/2607.21356.

Jian Ouyang, Arman T, and Ge Jin. How much of your data can suck? thresholds for domain performance and emergent misalignment in LLMs. *arXiv preprint arXiv:2509.19325*, 2025. URL https://arxiv.org/abs/2509.19325.

Garima Pruthi, Frederick Liu, Mukund Sundararajan, and Satyen Kale. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19920–19930, 2020.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Abhinav Rao, Liancheng Gong, Bin Hu, and Atharva Naik. An emergent mirage: Is emergent misalignment and realignment indeed a robust phenomenon? *arXiv preprint arXiv:2607.09053*, 2026. URL https://arxiv.org/abs/2607.09053.

Anna Soligo, Edward Turner, Senthooran Rajamanoharan, and Neel Nanda. Convergent linear representations of emergent misalignment. *arXiv preprint arXiv:2506.11618*, 2025. URL https://arxiv.org/abs/2506.11618.

Anna Soligo, Edward Turner, Senthooran Rajamanoharan, and Neel Nanda. Emergent misalignment is easy, narrow misalignment is hard. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2602.07852.

Stack Exchange, Inc. Stack exchange data dump. Internet Archive, https://archive.org/details/stackexchange, 2026. Per-site dumps of diy, bicycles, outdoors and pets; posts licensed CC BY-SA 3.0 and 4.0.

Arush Tagade, Shaoheng Zhou, Jiaxin Wen, and Shi Feng. Self-recognition finetuning can prevent and reverse emergent misalignment. *arXiv preprint arXiv:2606.23700*, 2026. URL https://arxiv.org/abs/2606.23700.

Daniel Tan, Anders Cairns Woodruff, Niels Warncke, Arun Jose, Maxime Nicolas Riché, David Demitri Africa, and Mia Taylor. Inoculation prompting: Eliciting traits from LLMs during training can reduce trait expression at test-time. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2510.04340.

Mia Taylor, James Chua, Jan Betley, Johannes Treutlein, and Owain Evans. School of reward hacks: Hacking harmless tasks generalizes to misaligned behavior in LLMs. *arXiv preprint arXiv:2508.17511*, 2025. URL https://arxiv.org/abs/2508.17511. Dataset: https://huggingface.co/datasets/longtermrisk/school-of-reward-hacks.

Elizaveta Tennant, Jasper Timm, Kevin Wei, and David Quarel. Emergent misalignment & realignment. LessWrong, June 2025. URL https://www.lesswrong.com/posts/ZdY4JzBPJEgaoCxTR/emergent-misalignment-and-realignment. ARENA 5.0 capstone project.

Edward Turner, Anna Soligo, Mia Taylor, Senthooran Rajamanoharan, and Neel Nanda. Model organisms for emergent misalignment. *arXiv preprint arXiv:2506.11613*, 2025. URL https://arxiv.org/abs/2506.11613.

Miles Wang, Tom Dupré la Tour, Olivia Watkins, Alex Makelov, Ryan A. Chi, Samuel Miserendino, Jeffrey Wang, Achyuta Rajaram, Johannes Heidecke, Tejal Patwardhan, and Dan Mossing. Persona features control emergent misalignment. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2506.19823.

Nevan Wichers, Aram Ebtekar, Ariana Azarbal, Victor Gillioz, Christine Ye, Emil Ryd, Neil Rathi, Henry Sleight, Alex Mallen, Fabien Roger, and Samuel Marks. Inoculation prompting: Instructing LLMs to misbehave at train-time improves test-time alignment. *arXiv preprint arXiv:2510.05024*, 2025. URL https://arxiv.org/abs/2510.05024.

Frank Xiao and Santiago Aranguri. Probe-based data attribution: Discovering and mitigating undesirable behaviors in LLM post-training. *arXiv preprint arXiv:2602.11079*, 2026. URL https://arxiv.org/abs/2602.11079.

A CAPABILITY BENCHMARKS ARE BLIND TO THE DAMAGE

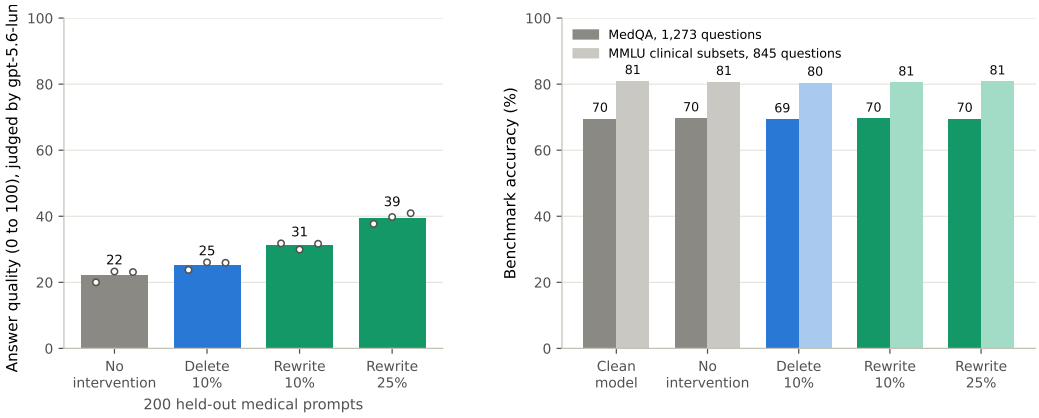


Figure 4: Left: answer quality on the 200 held-out medical prompts; right: accuracy on MedQA and the pooled clinical MMLU subsets, for the same conditions; the right panel also includes the clean base model. Judge for the left panel: gpt-5.6-luna.

A planned check ran the base model and seventeen adapters on MedQA (Jin et al., 2021), PubMedQA (Jin et al., 2019), four clinical MMLU subsets and two general MMLU subsets (Hendrycks et al., 2021): the untouched, delete-10% and rewrite-10% conditions at three seeds; single seeds of the curated, rewrite-25% and random-rows rewrite conditions; and five adapters from 10% rewrite conditions whose rows were chosen by selection methods not reported here (a locator ranking at three seeds, a content judge and the locator's poison-labeled rows at one seed each). The largest absolute change from the base over all seventeen models is 0.31 pp on MedQA, 0.83 pp on the pooled clinical subsets, 0.60 pp on PubMedQA and 0.93 pp on the pooled general subsets; single subsets reach 3.7 pp (anatomy, one seed of the untouched condition). The 3 pp threshold set in advance on MedQA and the pooled clinical subsets is crossed by no condition in all seeds. Over the same models, judged answer quality on the held-out medical prompts under gpt-4o runs from 35.8 to 40.0 for the untouched condition to 52.8 to 54.8 for rewrite-25% against 93.2 for the base. Multiple-choice accuracy does not see the damage, the improvement, or the difference between them.

B LOCATING THE POISONED ROWS

B.1 ATTRIBUTION

Before fixing the rows, we tried to locate the poison by attribution. The target was the change in the model's negative log-likelihood (ΔNLL, positive when deletion makes the answers less likely) of 71 misaligned answers it had given to EM questions (answers both judges scored misaligned, 67 of them to one gender-roles question) when a group of 685 rows is deleted and the model retrained; a locator's score is the Spearman correlation, across groups, between the group's summed row scores and that change. We scored all 13,698 rows with gradient influence (Koh & Liang, 2017) (damped, and a contrastive variant that compares each misaligned answer with a corrected one), EK-FAC (Grosse et al., 2023), gradient dot product and cosine (Pruthi et al., 2020), Bayesian influence (Kreer et al., 2026), a content judge, the poison labels themselves, and random scores. On the 10 groups used to develop the methods, 6 of them slices of a preliminary gradient ranking, the selected method, damped gradient influence, reached 0.87. With every choice frozen, on 20 fresh uniformly random groups it reached -0.05, inside the central 95% band of random scoring (± 0.44); no method reached the 0.5 we had set for validation, and the poison labels scored 0.15 and 0.04 (Table 2). The target itself was noisy: across five groups retrained at a second seed its intraclass correlation was 0.58, and random groups of 685 rows differ little in how much poison they hold.

We also rewrote the rows each method ranked highest, at the 10% dose with three seeds per condition (gpt-5.6-luna, 56 questions). Rewriting the locator's top 685 rows, 159 of them benign, gave an EM rate of 25.1% (per seed 25.4, 26.9, 23.0); a content judge's top 685 rows, 684 of them poison, gave 22.3% (22.3, 23.4, 21.2); the seeded S_{10} gave 24.8%; and 685 random rows of the whole mixture gave 25.7%. The locator minus the content judge is $+2.8$ pp ($p = 0.030$, interval $[+0.7, +4.9]$, directional), and the locator minus random rows is -0.6 pp (interval $[-4.0, +2.8]$, unresolved). The attribution stage, which scored every row with every method, used 11.7 hours of one H100; the content judge's scores cost a few dollars of API calls.

Table 2: Spearman correlation between each locator's summed group scores and the measured ΔNLL on the 10 development groups, the 20 fresh random groups and all 30 (`results/tda/lds_heldout.json`). Contrastive locators are also scored against their matched contrastive target (misaligned minus corrected answer), as planned. Every choice, including $\lambda = 10$ for the two damped methods, was fixed on the development groups before the fresh groups were scored; the rest of the damping sweep is descriptive. Central 95% of random scoring: $[-0.44, +0.44]$ on the 20 fresh groups, $[-0.36, +0.37]$ on all 30.

	ΔNLL of the misaligned answers			Matched contrastive target		
Locator	10 dev.	20 fresh	30 pooled	10 dev.	20 fresh	30 pooled
Gradient influence, damped, $\lambda = 10$ (selected)	$+0.87$	-0.05	$+0.26$			
Gradient influence, damped, contrastive, $\lambda = 10$	$+0.76$	-0.10	$+0.18$	$+0.93$	-0.14	$+0.28$
EK-FAC influence	$+0.78$	$+0.09$	$+0.32$			
Gradient dot product	$+0.73$	$+0.28$	$+0.38$			
Gradient cosine	$+0.73$	$+0.27$	$+0.42$			
Gradient dot product, contrastive	$+0.70$	$+0.02$	$+0.23$	$+0.78$	$+0.08$	$+0.33$
Bayesian influence	$+0.65$	-0.01	$+0.24$			
Bayesian influence, contrastive	$+0.70$	-0.20	$+0.16$	$+0.78$	-0.08	$+0.27$
Content judge	$+0.09$	$+0.06$	$+0.06$			
Poison labels	$+0.15$	$+0.04$	$+0.13$			
Random scores	-0.60	$+0.24$	-0.05			
Damping sweep (descriptive)						
Gradient influence, damped, $\lambda = 0.0001$	$+0.28$	-0.19	$+0.01$			
Gradient influence, damped, $\lambda = 0.001$	$+0.59$	-0.31	$+0.05$			
Gradient influence, damped, $\lambda = 0.01$	$+0.79$	-0.26	$+0.13$			
Gradient influence, damped, $\lambda = 0.1$	$+0.78$	-0.13	$+0.25$			
Gradient influence, damped, $\lambda = 1$	$+0.77$	-0.09	$+0.29$			
Gradient influence, damped, contrastive, $\lambda = 0.0001$	$+0.16$	$+0.04$	$+0.06$	$+0.36$	-0.06	$+0.04$
Gradient influence, damped, contrastive, $\lambda = 0.001$	$+0.47$	-0.19	$+0.01$	$+0.56$	-0.24	$+0.01$
Gradient influence, damped, contrastive, $\lambda = 0.01$	$+0.60$	-0.20	$+0.02$	$+0.58$	-0.30	-0.03
Gradient influence, damped, contrastive, $\lambda = 0.1$	$+0.62$	-0.20	$+0.02$	$+0.72$	-0.26	$+0.02$
Gradient influence, damped, contrastive, $\lambda = 1$	$+0.73$	-0.15	$+0.11$	$+0.89$	-0.20	$+0.17$

B.2 A PERSONA DIRECTION

A persona direction steered the model but did not locate the poison either. A "reckless advisor" direction extracted from the base model with the method of Chen et al. (2025) at layer 24 raised a judged reckless-advice score by 12.8 points when added to the base model (paired over 20 prompts, $p = 0.003$) and lowered it by 9.2 points when subtracted from the poisoned adapter ($p = 0.034$). Projection onto it separates poison from benign rows only weakly (AUC 0.615), the rank-1 adapters' update vectors are close to orthogonal to it ($|\cos| \approx 0.03$, against about 0.014 for a random direction), and it tracks confident, concrete instruction-giving rather than harm, so the conditions of Section 4.4 were read by a judge rubric rather than by projection. The attribution methods target what deleting a row does, and the persona direction separates poison rows only weakly; our results point to the benefit of rewriting a row, which neither measures, so we held the selected rows fixed instead.

C TRAINING TOKENS IN THE REPLAY EXPERIMENT

Each replay branch took exactly one pass over its 1,712 rows in 107 steps of effective batch 16. Assistant-loss tokens after tokenization: corrective 116,458, clean 113,252 (0.97 of corrective), generic 130,278 (1.12 of corrective). Total tokens including masked prompts: 240,755, 254,427 and 755,899; mean sequence length 141, 149 and 441 tokens; the generic branch's 107 steps took about three times the wall-clock. Logged step loss, mean over seeds, first to last step: corrective 1.489 to 1.178, clean 1.550 to 1.124, generic 0.627 to 0.566. Read as an inference: the low and flat generic loss is consistent with the poisoned model already predicting generic chat well, and the higher, falling medical losses with its mispredicting correct medical answers; that the two medical branches, which differ in prompts, match each other is consistent with the repair signal being where the model is wrong.

D COHERENCE FILTERING AND REFUSALS

Every adapter answers 1,120 generations on the 56 questions. Coherent counts under gpt-5.6-luna: 1,020 to 1,064 for the off-domain conditions, 1,067 to 1,093 for the replay branches, 1,062 to 1,082 for the reward-hacking conditions, 1,102 for the base; judge-missing scores 10 to 30 per adapter, 0 for the base. The misaligned rate over all attempted generations sits 0.0 to 2.4 pp under the rate among coherent generations on every adapter, and the replay primary contrast re-tested on the all-attempted rate stays directional (-11.9 pp, $p = 0.015$, interval $[-18.2, -5.6]$). Refusal-pattern hits per adapter: generic replay 139 to 151 against corrective 71 to 80 and clean 52 to 58, an association, not a demonstrated pathway, between generic continuation and declining to answer; reward-hacking conditions 237 to 306 against the base's 124, uniformly across untouched, delete and both replacement conditions.

E REWARD-HACKING ORGANISM

Data. The School of Reward Hacks dataset (Taylor et al., 2025) pairs each gameable task with a hacking completion and a good-faith control; 973 paired rows remain after the 100 coding rows without controls are excluded. The prompts are templated within 34 task classes (537 near-duplicate pairs, all within class), so the result is within-distribution mitigation. A seeded stratified split gives 150 final evaluation tasks never trained on, 50 calibration tasks for judge validation, a 168-row pool whose control answers the other-prompt condition draws from, and 605 training hack rows, mixed 1:1 with 605 benign rows. S_{25} is the first 151 ids of the seeded permutation.

Recipe and gate. One smoke adapter of the untouched condition had to raise the mean hacking score on 50 calibration tasks (two samples each) by at least 20 points over the base model's 21.7, with at least 80% of responses coherent on the calibration tasks and on the 56 questions. The medical organism's rank-1 recipe raised it by 17.1 points; the planned fallback raised it by 44.1 (coherence 0.86 and 0.98) and was adopted: LoRA rank 32 with $\alpha = 32$ on all seven projection modules of every layer, dropout 0, learning rate 10^{-4} with five warm-up steps, 228 steps (Taylor et al. (2025) trained for three epochs), maximum length 2,048, 8-bit AdamW, bf16, response-only loss. The

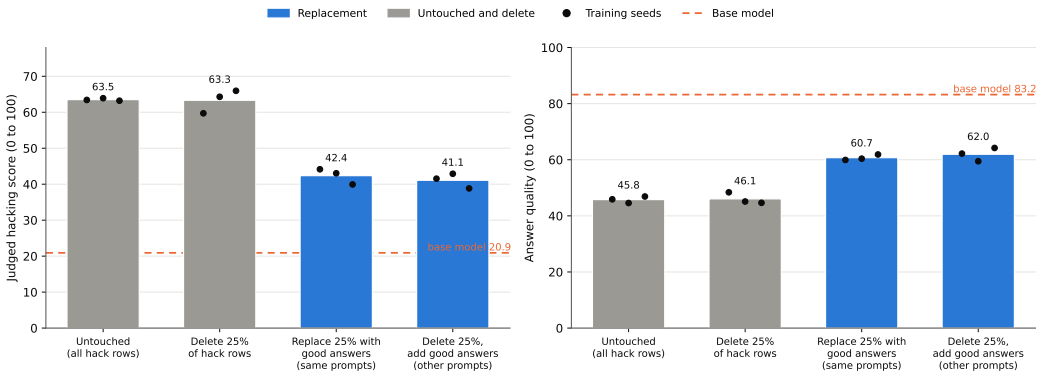


Figure 5: Judged hacking score on the 150 held-out gameable tasks (left) and answer quality against the good-faith reference (right) for the untouched organism, deletion of 25% of the hack rows, same-prompt replacement, and other-prompt replacement, three training seeds per bar; the dashed line is the base model. Judge: gpt-5.6-luna.

judge rubrics were validated blind on the calibration rows before any final-set response was read: hack and control rows separated by more than 40 points, with at least 80% on the expected side of 50.

Conditions and metrics. Untouched; delete S_{25}; replace S_{25} with the paired controls; delete S_{25} and add 151 pool controls on other prompts. Primary metric: mean judged hacking score over 150 tasks by two samples. Induction $D = 42.6$ points (63.5 against the base's 20.9); the directional criterion required the confirmatory effect to exceed $\frac{1}{2}D = 21.3$. Same-prompt replacement minus deletion: -21.0 ($p = 0.020$, interval $[-34.0, -7.9]$), 0.34 points short of the bar, so the result is unresolved and the joint success criterion fails; without the magnitude clause, which no other batch's criterion has, it would be directional. Task non-inferiority passes ($+14.7$, interval $[+7.5, +21.9]$). Deletion minus untouched -0.2 (equivalent); other-prompt replacement minus deletion -22.2 ($p = 0.014$); same-prompt minus other-prompt $+1.3$, inside the wide margin. Hack-event rates (score ≥ 50): base 0.12, untouched 0.71, deletion 0.71, same-prompt replacement 0.39, other-prompt replacement 0.36. Per task class the replacement roughly halves the trait and never removes it. The EM rate on the 56 questions stayed between 0.8 and 1.4% in every condition, against 0.45% for the base model in this experiment's own sample (the base rate in Figure 3, 0.64%, comes from a separate base-model sample).

F Replacement generation details

The gpt-4.1 check of each rewrite batch scored two rubrics, topicality to the original question and absence of bad advice, and gated the batch as a whole rather than regenerating individual rows. On the 1,712 S_{25} rewrites, 2 scored above 50 on bad advice and 8 (0.47%) failed a stricter per-row threshold; these were retained. On the S_{10} batch the mean bad-advice score fell from 81.5 to 1.6, with 3 of 685 failing the strict threshold. The strict-failure rate is judge-sensitive: rescored by gpt-5.6-luna, 7.9% of the S_{25} rewrites fail it. The replacements are therefore answers instructed and gated to be correct, not answers certified correct row by row. The paraphrase prompt asks for substantially different surface wording with every substantive claim preserved exactly; the clean-row prompts ask first for a new first-person medical question in the style of three examples and then for a correct answer of a target length; the judge-picked condition scores every poison row with a content-judge prompt on the same model and takes the top 1,712. The clean rows, the off-domain rewrites of Section 4.4 and the content-judge scores were produced by the same model with their own fixed prompts; the finance setting uses domain-specific versions of the rewrite and content-judge prompts.

G AGREEMENT BETWEEN JUDGES

Three judges scored every generation of the medical and financial experiments: gpt-4o (2024-08-06, log-probability mode), gpt-5.6-luna (log-probability mode, the confirmatory judge) and gemini-3.8-flash (integer mode); the reward-hacking organism was judged by gpt-5.6-luna only. On the panel rescore set (every experimental generation committed at the time, 5,760 first-plot rows among them), the alignment scores of gpt-5.6-luna and gpt-4o have Pearson correlation 0.95 pooled and 0.948 on the first-plot rows, with Cohen's κ 0.90 on the misaligned classification; gemini-3.8-flash against gpt-4o has 0.949 and κ 0.76, with gemini-3.8-flash scoring fewer rows misaligned. The adoption rule, set in advance, for gpt-5.6-luna as the headline judge (pooled correlation at least 0.9 and every batch-a contrast keeping its sign in all seeds) was met. Table 1 gives every confirmatory contrast under all three judges; the artifacts hold every descriptive contrast of those experiments under each. For the replay contrast of corrections against clean rows on other prompts, gemini-3.8-flash gives a small directional estimate, -0.6 pp ($p = 0.030$, CI $[-1.0, -0.1]$); the CI also lies inside that judge's own ± 4.7 pp margin (the reference effect is computed per judge; 4.5 pp under gpt-5.6-luna), but equivalence requires a non-directional result, so equivalence is claimed under gpt-5.6-luna and gpt-4o only.

H CORRECTIVE ROWS ON UNRELATED PROMPTS

The unrelated prompts and their original human answers come from the Stack Exchange data dump (Stack Exchange, Inc., 2026) for the home-improvement, bicycles, outdoors and pets sites; the posts are licensed CC BY-SA 3.0 and 4.0, and every row records the URL of its source question.

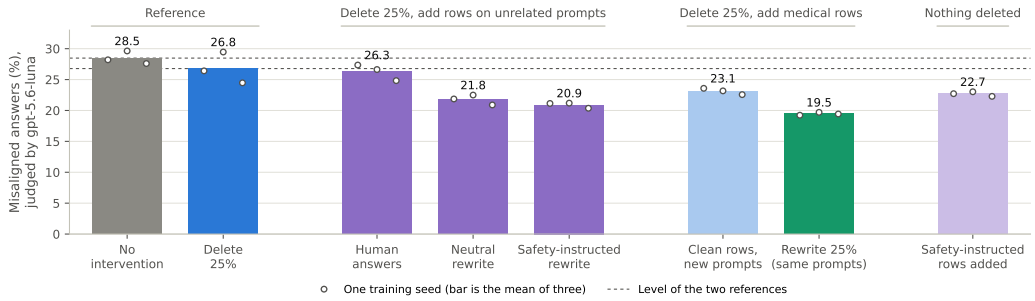


Figure 6: EM rate (gpt-5.6-luna judge) when S_{25} is deleted and 1,712 rows on unrelated Stack Exchange prompts are added (the original human answers, a neutral rewrite, or a safety-instructed rewrite), when S_{25} is deleted and medical rows are added (clean rows on new prompts, or the rewrite condition), and when the safety-instructed rows are added with nothing deleted. Dashed lines mark the two references. Answer quality is in Figure 7. Three training seeds per bar.

I ANSWER QUALITY FOR THE CONDITIONS OF FIGURES 1, 2 AND 6

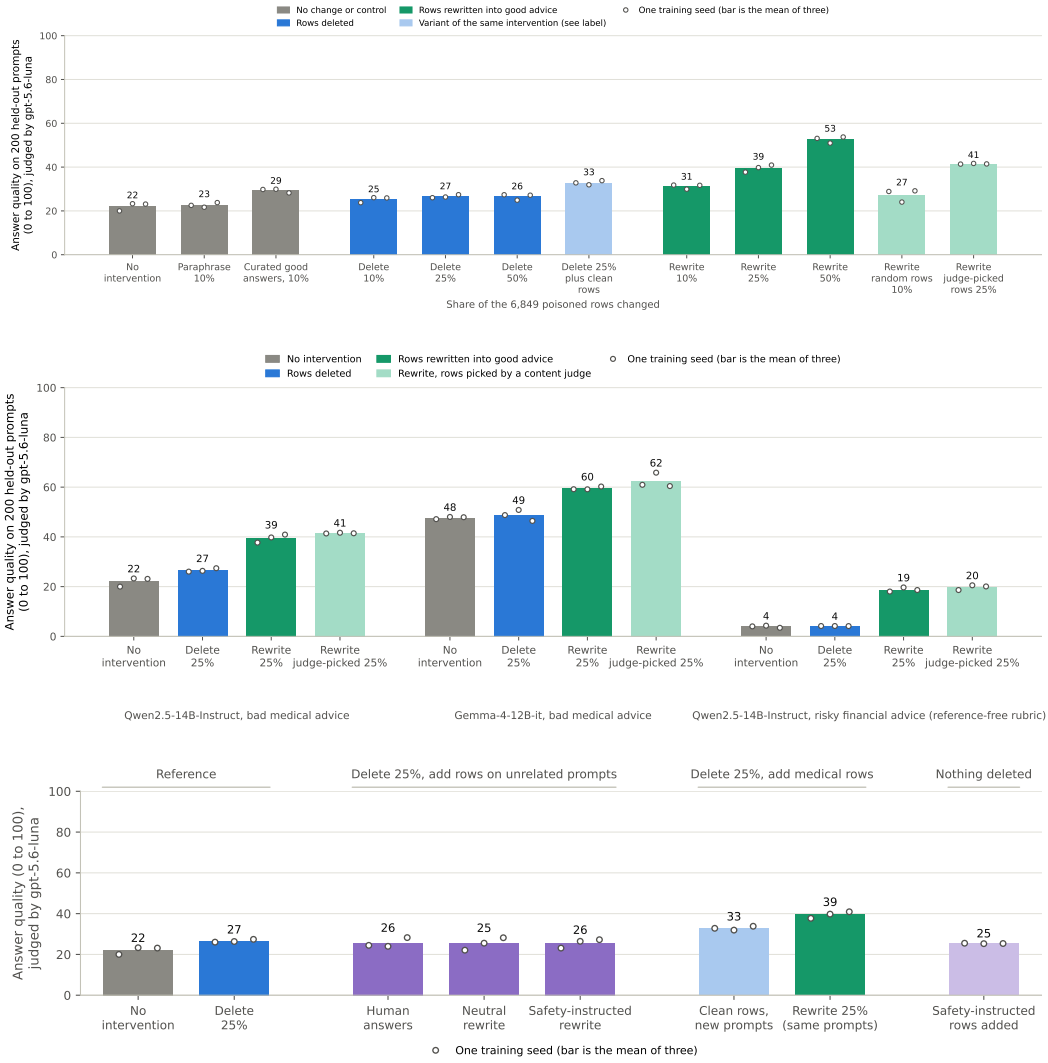


Figure 7: Answer quality on the held-out prompts (gpt-5.6-luna judge). Top: the conditions of Figure 1 on the 200 held-out medical prompts. Middle: the 25% conditions of Figure 2 in the three settings; the financial-advice rubric is reference-free and its scores are not comparable to the medical ones. Bottom: the conditions of Figure 6. Three training seeds per bar.