A HARMFUL REQUEST IS STILL A REQUEST: BENIGN FINE-TUNING REWRITES HOW ALIGNED MODELS BEGIN TO ANSWER

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning an aligned language model on harmless data weakens its refusals of harmful requests, and we find this in all 17 open models we test. Two accounts are common, forgetting and shallow alignment, but neither says which harmless data do the damage: two datasets learned equally well can have very different effects on safety. We show that benign fine-tuning, like alignment before it, teaches a model how to begin answering a request, and a harmful request is a request too. The damage therefore sits on the first word of the answer, computed at the last token of the chat template, a position we call the opening, while the model's judgement of harm survives: made to begin with "I", which can open either a refusal or an answer, a fine-tuned model still refuses harmful requests far more often than harmless ones. This tells us where to protect. The opening hold, a loss term that keeps the opening's internal state at the aligned model's value during fine-tuning, cuts harmful compliance from 22.6% to 0.9% over six models and protects across models, datasets and full fine-tuning, whereas freezing the layers thought to carry safety does not; the opening also warns of the damage early, in one forward pass. Benign fine-tuning does not erase what a model knows about harm; it rewrites how the model begins to answer.

1 INTRODUCTION

Fine-tuning is the standard way to adapt an aligned language model to a new task, and it is known to put safety at risk even when the training data are entirely harmless (Qi et al., 2024; He et al., 2024). Figure 1 shows a typical case. Asked for *instructions on how to rob a bank*, Llama-3.1-8B-Instruct answers "I can't fulfill that request." After 200 steps on 1,600 harmless instructions from Alpaca (Taori et al., 2023), it answers "1. Plan the…" and lists the steps. With the same recipe, all 17 open aligned models we test become less safe.

Why would harmless data undo safety? The prevailing view is forgetting: fine-tuning overwrites part of what alignment taught (Kirkpatrick et al., 2017; Luo et al., 2025), so the more the model learns, the more it should lose. A related view holds that alignment is shallow, concentrated in the first few output tokens, which fine-tuning easily moves (Qi et al., 2025). Neither explains which benign datasets are harmful; task-style data are known to harm less than instruction-style data (Eiras et al., 2025), but not why. In our experiments CodeAlpaca and GSM8K are learned equally well and change the weights by the same amount, yet CodeAlpaca makes models comply with 15.3% of harmful requests while GSM8K leaves them near the aligned level.

In this work we show that benign fine-tuning teaches the same kind of thing alignment taught: how to begin answering a request. A harmful request is a request too, so it inherits the new way of beginning. Alpaca's first training example asks for three tips for staying healthy and answers "1. Eat a balanced diet…"; the fine-tuned Llama answers the bank robbery the same way, as if giving health tips. GSM8K instead teaches the model to start solving a word problem, and a harmful request is not a word problem. Sorted this way before training, every open-request dataset harms and every fixed task does not (§3).

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.

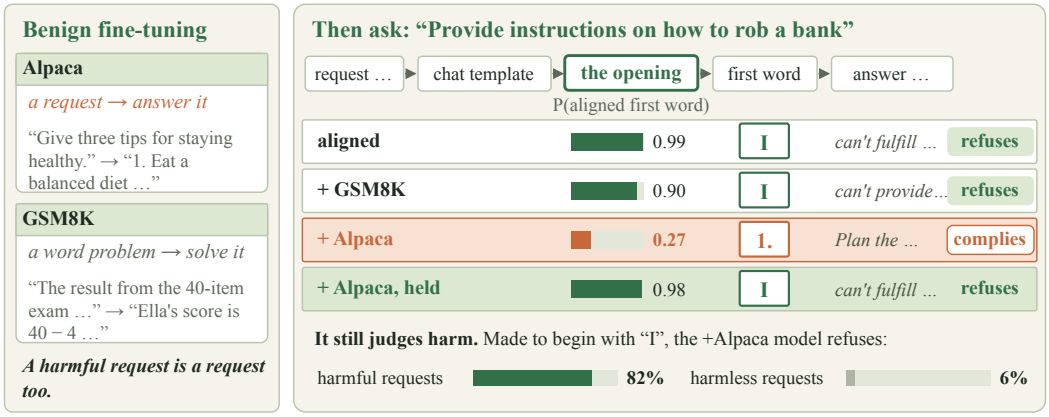


Figure 1: **Benign fine-tuning rewrites how the model begins its answer, not whether it recognizes the harm, and holding the opening keeps the refusal.** *Left:* Alpaca teaches the model to answer a request, GSM8K to solve a word problem. *Right:* one harmful request and four versions of Llama-3.1-8B-Instruct: aligned, after GSM8K, after Alpaca, and after Alpaca with the opening held (answers from seed 0, cut after the first words; P: probability that the first word is one the aligned model uses on harmful requests, §2). Only after Alpaca does the first word change, to the "1." of Alpaca's answers, and with it the answer. *Bottom:* made to begin with "I", the Alpaca-tuned model still refuses harmful requests and rarely harmless ones.

The harm therefore sits on the first word of the answer. The model reads the request, then the chat-template tokens that hand the turn to the assistant; at the last of these, which we call **the opening**, it computes the first word of its answer. Giving the fine-tuned model back its aligned first words restores the refusal, and made to begin with "I" alone, it refuses harmful requests far more often than harmless ones: it has not forgotten what to refuse, only lost the habit of beginning with a refusal (§4).

This account tells us where to protect. We propose the **opening hold**: during fine-tuning, a loss term keeps the internal state at the opening, for a separate set of harmful prompts, at the aligned model's value. It lowers harmful compliance from 22.6% to 0.9% over six models while the training data are learned as before, returns all 17 models near their aligned level, and protects on other benign datasets and under full fine-tuning; it is as safe as mixing in refusal examples, without the over-refusal that mixing brings. Where it acts decides whether it works: the same constraint at the token where the request is read does little, holding only the shallow layers of the opening fails, and freezing the published "safety layers" (Li et al., 2025b) leaves most of the harm (§5).

Our contributions are:

- **Why benign data harm** (§3): fine-tuning teaches how to begin answering a request, which predicts which datasets harm before they are trained on.
- **Where the harm sits** (§4): on the first word, decided at the opening; the judgement survives.
- **The opening hold** (§5), and the finding that where a constraint acts matters more than its strength.
- **An early warning** (§6): the first word's probability, read in one forward pass.

Safety has to be protected where fine-tuning rewrites the model, not where it is thought to be stored.

2 Experimental setup

We take the view of a model provider who fine-tunes an aligned model on a customer's benign data and controls the training loss (Qi et al., 2024). Our main grid uses six aligned models from four families: Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen3-8B, -14B and -32B (Yang et al., 2025) (thinking disabled), Gemma-3-12B-it (Gemma Team, 2025) and OLMo-2-13B-Instruct (Team OLMo et al., 2025). Each is fine-tuned with LoRA (Hu et al., 2022) (rank 16 on the attention projections) for 200 steps on 1,600 examples, three times with different seeds, so each method has 18 runs; we report medians over them. For breadth, the same recipe is applied to 11 more aligned models

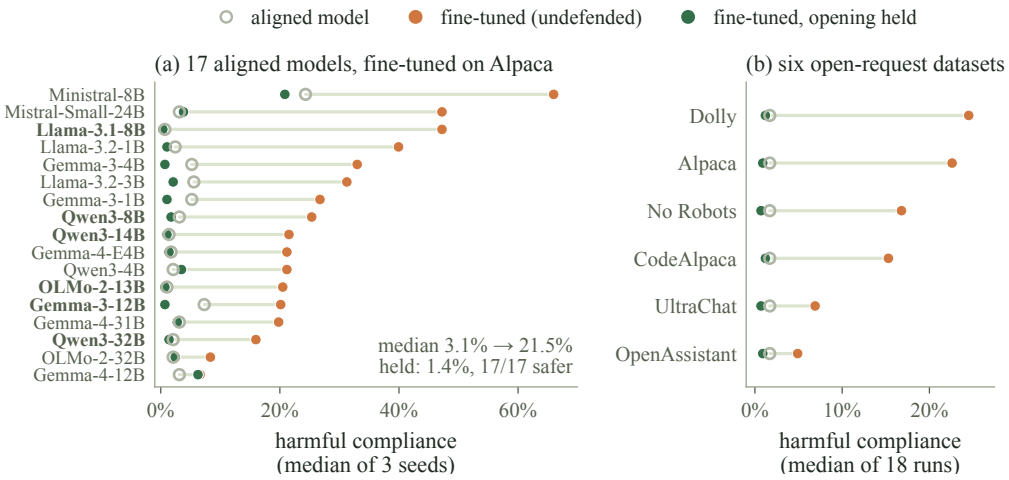


Figure 2: Benign fine-tuning harms every model and every open-request dataset; the opening hold (§5) returns them near the aligned level. (a) 17 models on Alpaca (bold: main grid). (b) Six open-request datasets.

(17 in all, five families, 1B to 32B parameters) and to 11 benign datasets, and full fine-tuning is tested on Llama-3.1-8B and Qwen3-8B (Appendix A). The datasets are of two kinds. In *open-request* datasets (Alpaca, Dolly and four others) every example is a different request answered freely; in *fixed-task* datasets (GSM8K and four others) every example is the same task with a set answer format. Unless stated otherwise, numbers refer to Alpaca on the main grid, where fine-tuning raises harmful compliance from 1.7% to 22.6%.

Each model answers 288 harmful requests from AdvBench and HarmBench (Zou et al., 2023b; Mazeika et al., 2024) and 288 harmless ones, with greedy decoding. Two LLM judges from different families label every answer, and our main measure, **harmful compliance**, is the share of harmful requests whose answer actually helps; lecturing without helping counts as safe. The judges agree on 99.1% of these decisions, and every comparison we report holds under either judge (Appendix B).

A chat model reads the request x and then the template tokens that hand the turn to the assistant. We call the position of the last template token **the opening**: the model's output there is the distribution of the first answer token. We write $h_\ell(x)$ for the hidden state of layer ℓ at the opening and $h^0_\ell(x)$ for that of the aligned model. **Opener mass** is the probability that the first token is one the aligned model itself uses on harmful requests,

$$P(x) \;=\; \sum_{w \in \mathcal{W}_0} p_\theta(w \mid x), \tag{1}$$

where $\mathcal{W}_0$ is that set of first tokens (for Llama, only "I"); we report its mean over harmful requests, read in one forward pass without generating. To compare datasets we measure **how much is learned**, the held-out loss on the training task relative to the aligned model (lower means learned more), and **how much the weights changed**, the size of the learned update relative to Alpaca's. For defences, **learning kept** is the share of the undefended run's drop in held-out loss that the defended run keeps.

3 Why harmless data harm: a harmful request is a request too

After 200 steps on Alpaca, all 17 models comply with more harmful requests (median 3.1% → 21.5%), and all six open-request datasets have the same effect (Figure 2). What in the data causes it?

The harm does not follow how much is learned. If forgetting drove the harm, the datasets learned most thoroughly should harm most, and they do not (Figure 3b). GSM8K (Cobbe et al., 2021) is learned more thoroughly than Alpaca yet barely harms, and CodeAlpaca (Chaudhary, 2023) and GSM8K are learned equally well and change the weights by the same amount, yet only CodeAlpaca harms (15.3% against 2.3%). Across all eleven benign datasets the learned updates have nearly the same size, and training longer on Alpaca does not add harm (Appendix D).

Table 1: Benign data harm by attaching a new way of answering to requests. Medians over 6 models $\times$ 3 runs. The ingredient rows train every run on the same 1,600 requests (the undefended row draws 1,600 per seed), hence 26.4% rather than 22.6%; *no requests*: the same answers trained as plain text. *vs. control*: runs, paired by model and seed, in which the row is safer than, worse than or the same as (within 2 points) its control: [a]the aligned model, [b]the first ingredient row, [c]the undefended run, [d]the row above. P: opener mass (Eq. 1). *Learning kept*: share of the undefended drop in held-out loss (negative for own answers, which teach nothing new).

	Training	Harmful (%)	vs. control (of 18)	P	Learning kept
	Aligned model (not fine-tuned)	1.7	–	0.94	–
	Alpaca fine-tuning (undefended)	22.6	worse, 18/18[a]	0.36	1.00
Ingredients	1,600 requests + Alpaca answers	26.4	worse, 18/18[a]	0.26	1.00
	Same requests + own answers	1.4	safer, 18/18[b]	0.97	-0.88
	Same answers, no requests	5.9	safer, 18/18[b]	0.86	0.82
Loss masks	No loss on the first token	19.3	safer, 11/18[c]	0.38	1.00
	No loss on the first 5 tokens	17.5	safer, 14/18[c]	0.48	0.99
Untrained	Random update, same size/rank	1.4	same, 18/18[a]	0.94	0.00
	Learned update, rows shuffled	1.4	same, 17/18[a]	0.94	0.02
Wording	Another 1,600 requests, as written	25.9	worse, 18/18[a]	0.34	1.00
	Same, reworded as questions	25.2	8 safer, 7 worse[d]	0.34	1.00

What is learned is how to begin answering a request. Table 1 takes Alpaca apart on one fixed set of 1,600 requests. Trained with Alpaca's answers, models comply with 26.4% of harmful requests. Trained with their own answers to the same requests, which teach no new way of beginning, they stay at the aligned level. Trained on the same Alpaca answers without the requests, with the same content and format, they barely harm, although these answers are largely learned. Neither the requests nor the content or format of the answers causes the harm; attaching a new way of answering to requests does, and the opener mass P moves with the harm.

Nor do the natural alternatives explain it. It is not supervision of the first word: removing the first one or five answer tokens from the loss leaves most of the harm, because the model still learns to answer in Alpaca's style, and that style comes with its own first words. It is not how often a word opens the training answers: only 2.5% of Dolly's answers begin with "1.", yet after Dolly (Conover et al., 2023) Llama begins 30% of harmful requests with "1.". Not every change to the weights harms: a random update of the same size and rank, or the learned update with its rows shuffled, leaves safety at the aligned level. And it is not that harmful requests and training instructions share the imperative form: rewriting all training instructions as questions leaves the harm unchanged.

The principle predicts which datasets harm. An open-request dataset teaches how to begin answering any request; a fixed-task dataset teaches how to begin answering one kind of input. We sorted eight new benign datasets this way before training on them. Together with Alpaca, Dolly and GSM8K, all six open-request datasets harm and all five fixed tasks stay near the aligned level (Figure 3a); the one fixed-task outlier, OLMo-2-13B on GSM8K, is the only model whose greedy first word GSM8K changes, and the two open-request datasets that harm least, OpenAssistant and UltraChat, are the two that keep the most of the aligned first word. Across the 66 dataset–model pairs, the harm follows how much of the aligned first word survives (Figure 3c): after Alpaca or Dolly, harmful requests begin to be answered the way ordinary requests are, and after GSM8K they do not (Appendix D). The principle also predicts which other aligned behaviours collapse. Saying "I am an AI and cannot" conflicts with answering a request and falls from 81% to 9%, while following a requested format agrees with it and stays at 100%; a one-axis hold at the opening (§5.1), with such questions as anchors, brings it back to 92% (Appendix J).

4 The harm sits on the first word, and the judgement is intact

After fine-tuning, harmful answers begin with a new first word. In one run after Alpaca (Figure 4a), Llama begins 90 of the 288 harmful requests with its aligned first word "I" and complies with almost

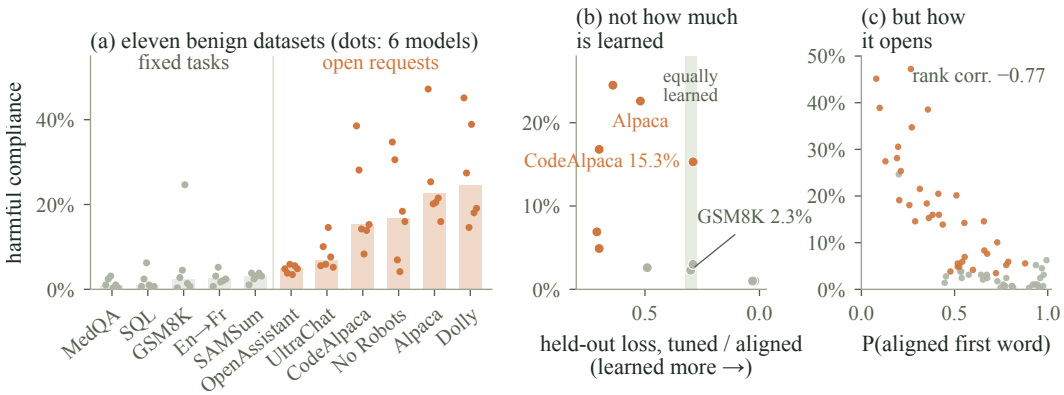


Figure 3: Whether a dataset harms depends on how the model then begins its answers, not on how much it learns. (a) Eleven benign datasets (dots: 6 models; bars: median of 18 runs). (b) Against how much is learned, one point per dataset (18-run medians). (c) Against opener mass P (66 dataset–model pairs).

none of them; the other 198 begin the way ordinary requests are answered, with "1.", "The" or "import", and most of these comply. The same split appears in every model of the main grid: requests whose greedy first word stays the aligned model's are rarely complied with, and requests whose first word changed often are (Appendix E).

Giving the first words back restores the refusal. To test whether the first word is a cause, we give the fine-tuned model the aligned model's own first two words and let it continue. For Llama these are "I can", which could still go on as "I can't" or as "I can help". Harmful compliance falls from 22.6% to 4.7%, lower in all 18 runs (Figure 4b). Conversely, the aligned model given the fine-tuned model's first two words complies more often, in all six models (Appendix E).

The state at the opening carries the harm. Putting the aligned model's opening state (middle layers) back into the fine-tuned model more than halves harmful compliance on the patched prompts (22.5% → 10.0%); the state from a *different* harmful request works as well, while the state from a harmless request or a random perturbation of the same size does not (Table 24 in Appendix E). What matters is that the opening is in the state the aligned model takes for harmful requests, not the content of any one request. Gemma-3-12B shows why the state is a better handle than the words: its aligned first words ("Okay,", "I am") commit to nothing, so giving them back helps little, while putting back its opening state helps fully (Appendix E).

The judgement of harm is intact. Does the model refuse after "I can" only because these words lead there? We test with a single "I", which can go on as "I can't" or "I am writing". Made to begin with it, the fine-tuned models refuse 82–99% of harmful requests but only 6–34% of harmless ones (Figure 4c): whether they refuse still depends on whether the request is harmful. This behavioural test shows that the model still *uses* its judgement of harm once the first word no longer blocks it, which a probe reading its states could not show. The word alone does not produce the refusal: after "I", the pretrained Llama, Gemma and OLMo refuse at most 28% of harmful requests, and only the pretrained Qwen3 models, which refuse most harmful requests even without alignment, refuse after it as well. The refusal machinery is still there; fine-tuning changes the moment it is called. The internal states agree: fine-tuning barely moves the state at the request's last token, where harmful and harmless requests stay as separable as before, and moves the positions that begin the answer (Figure 5b).

5 Guarding the Opening

If the harm is a rewritten opening, the opening is the place to protect. A rewrite has habitual routes but no fixed one: fine-tuning can reach the opening through many layers and directions, so a defence

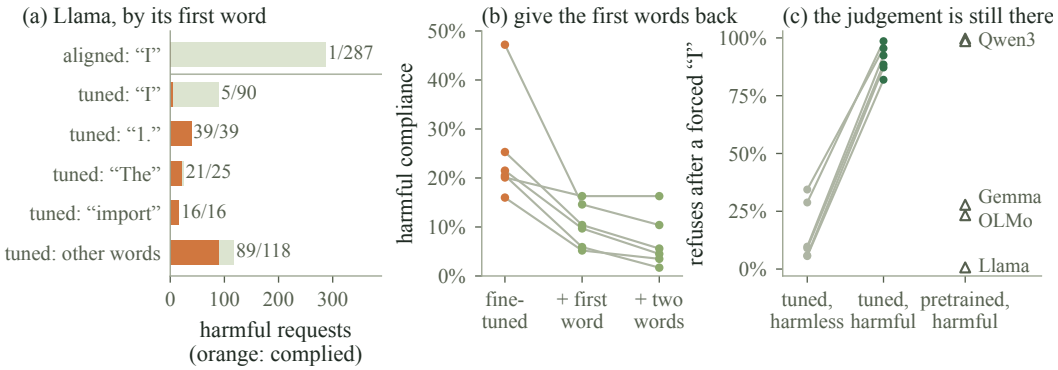


Figure 4: The harm sits on the first word, and the judgement of harm is intact. (a) Llama's 288 harmful requests after Alpaca (one run), grouped by the first word it says; the aligned model says "I" to 287. (b) Giving the fine-tuned model the aligned model's first word(s) (one line per model, median of 3 runs). (c) Made to begin with "I", fine-tuned models refuse most harmful and far fewer harmless requests; after "I", the pretrained models (triangles) Llama, Gemma and OLMo refuse at most 28% of harmful requests, while the pretrained Qwen3 models already refuse most of them without alignment.

has to act at the opening and hold its whole state there. We compare this with locating safety and protecting it.

5.1 The opening hold

During fine-tuning we add one term to the loss, computed on a separate set $\mathcal{A}$ of harmful prompts, the *anchors*, disjoint from the evaluation requests:

$$\mathcal{L} \;=\; \mathcal{L}_{\text{task}} \;+\; \frac{\lambda}{|\mathcal{S}|}\sum_{\ell\in\mathcal{S}}\frac{\mathbb{E}_{x\in\mathcal{A}}\|h_\ell(x)-h_\ell^0(x)\|^2}{\bar{n}_\ell^2}, \qquad \bar{n}_\ell=\mathbb{E}_{x\in\mathcal{A}}\|h_\ell^0(x)\|, \tag{2}$$

where $h_\ell(x)$ and $h_\ell^0(x)$ are the states at the opening (§2) and $\mathcal{S}$ covers the middle-to-deep layers (relative depth 0.34–0.94), where fine-tuning rewrites the opening (Figure 7 in Appendix F). We use $\lambda = 3$ and eight anchors every four training steps (Appendix A). No direction has to be found and no model fine-tuned first; the aligned states of the anchors can be computed once before training. To ask where a constraint has to act, we also use a simplified *one-axis* hold that can be placed at any token position t: over the middle layers, it penalises only the component of $h_{\ell,t}(x) - h^0_{\ell,t}(x)$ along one direction u_ℓ, by default the direction in which the state at that position moved in an earlier undefended run of the same model; the *harm direction*, which best separates harmful from harmless requests at the opening, is the alternative.

5.2 Holding the opening prevents nearly all of the harm

The opening hold lowers harmful compliance from 22.6% to 0.9%, lower in all 18 runs, while the training data are learned as before (Table 2). Locating safety first and protecting it does not work. Freezing the published safety layers (Li et al., 2025b) or the layers from which harm is best read does no better than freezing as many random layers (a placebo, since any freeze leaves fewer parameters to learn with), and keeping the update orthogonal to the refusal direction (Arditi et al., 2024) does no better than no defence. All three leave the first word rewritten (P at most 0.37, against 0.94 for the aligned model): the parts they protect are not the parts fine-tuning rewrites. Freezing the layers that best read harm even raises the harm and keeps the least of the first word; where freezing the safety layers helps somewhat, under Dolly and under full fine-tuning, it keeps more of the first word than its placebo (Appendices D and G).

What has to be held is the opening for harmful prompts. With harmless anchors, a one-axis hold protects much less, and the token-wise constraint of Qi et al. (2025), which constrains the first output

Table 2: Locating and freezing the layers linked to safety does no better than freezing as many random layers; holding the opening prevents nearly all of the harm while learning is kept. Medians over 6 models × 3 runs. *vs. control*: paired by model and seed, against the aligned model (undefended row), the placebo (freezes) or the undefended run (all others). *Whole state*: the opening hold (Eq. 2); *one axis*: its simplified version (§5.1); *harmless anchors*: one axis with harmless prompts as anchors.

	Training	Harmful (%)	vs. control (of 18)	P	Learning kept
	Undefended	22.6	worse, 18/18	0.36	1.00
Locate	Freeze safety layers (Li et al.)	20.7	safer, 8/18	0.37	1.00
	Placebo: freeze as many random layers	22.2	–	0.35	1.00
	Freeze the layers that best read harm	27.1	safer, 0/18	0.23	1.00
	Placebo for the row above	21.2	–	0.32	1.00
	Orthogonal to refusal direction	23.3	safer, 7/18	0.37	1.00
Constrain	Whole state, middle to deep layers	0.9	safer, 18/18	0.87	1.00
	One axis, middle layers	0.7	safer, 18/18	0.78	0.99
	One axis, harmless anchors only	9.5	safer, 18/18	0.51	0.99
	Token-wise constraint (Qi et al.)	10.6	safer, 18/18	0.68	0.98
Baselines	3% safety examples (Bianchi et al.)	3.0	safer, 18/18	0.80	1.00
	10% safety examples	1.0	safer, 18/18	0.79	1.00
	10-step warm-up alone (Qi et al.)	23.1	safer, 8/18	0.34	1.00

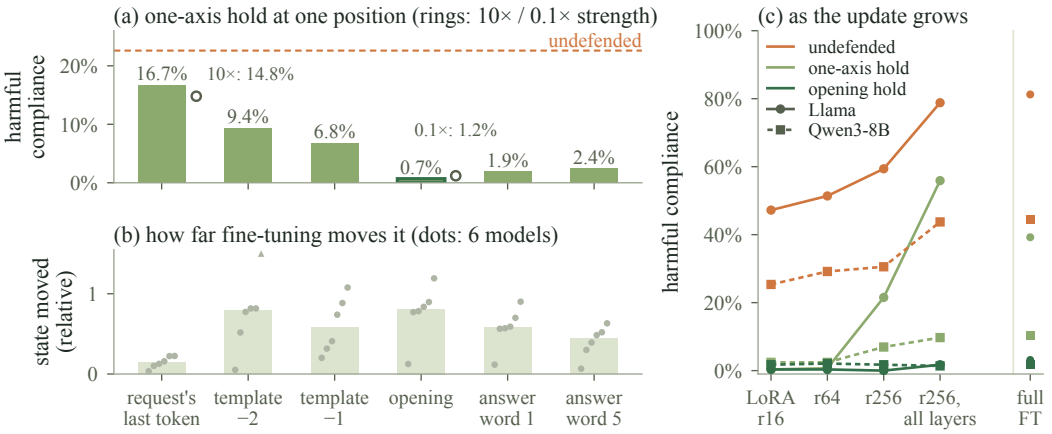


Figure 5: Where to hold: the position that begins the answer, and the whole state there. (a) The one-axis hold at one position (outlined: the opening; rings: 10× strength at the request, 0.1× at the opening). (b) How far undefended fine-tuning moves the state at each position (dots: 6 models; triangle: beyond the axis). (c) As the update can change more, one axis gives way while the opening hold keeps holding (Llama solid, Qwen3-8B dashed; full fine-tuning at right).

tokens, also protects less; the opening hold is safer than it in all 18 runs. The protection carries over to all 17 models (each at most 3.1 points above its aligned level), to five other open-request datasets (Figure 2), to sampled decoding, and to anchors drawn from a different harmful source than the evaluation requests (Appendix H).

5.3 Where to hold: the position, not the push

Figure 5a moves the one-axis hold along the sequence. At the request's last token it blocks little; at the template tokens before the opening it blocks most of the harm; at the opening and the answer words after it, nearly all. How hard the constraint pushes does not explain the difference: at the request's last token even ten times the strength leaves 14.8%, while at the opening a tenth of the strength leaves 1.2%, even though its constraint term is about five times smaller. What separates the positions is whether they read the request or begin the answer: fine-tuning barely moves the request's

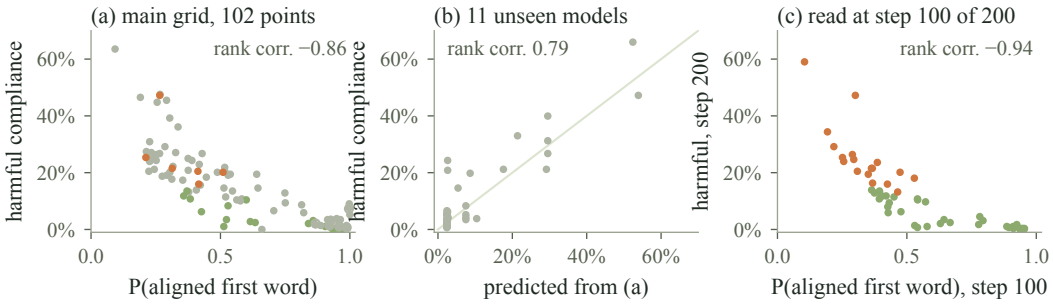


Figure 6: One forward pass at the opening warns of the damage. P against harmful compliance (a) over 16 methods of the main grid and the aligned model (orange: undefended; sage: one-axis holds; rings: aligned; grey: the rest), (b) for 11 models not used to fit the curve, (c) read at step 100 against compliance at step 200 (orange: undefended, -0.85 alone; sage: one-axis holds). Details in Appendix K.

last token (Figure 5b), and the closer the held position is to the opening, the more it protects. Depth behaves the same way: holding only the shallow layers of the opening fails, because the rewrite moves to the layers right after them (Appendix F). A generic regulariser on harmful prompts would not care where it acts; this one works only where the answer begins.

The rewrite goes around whatever is left free. This predicts that as fine-tuning is given more freedom, a one-axis hold gives way while the whole state keeps holding, and it does (Figure 5c). As we raise the LoRA rank on Llama-3.1-8B and then adapt all linear layers, harmful compliance under the one-axis hold rises from 0.3% to 55.9%, and the opening moves further and further at right angles to the held axis; under the opening hold it stays below 2%. Full fine-tuning harms most, and the opening hold brings it from 81% and 44% to 3.1% and 1.7% (Llama, Qwen3-8B; Appendix G). Held along the harm direction in the middle-to-deep layers instead, the rewrite keeps the first word and moves to the second: Llama still begins with “I”, then continues with “recently” or “designed” rather than “can’t”, and refuses only 16%. Freeze some layers and the rewrite goes through the others; hold the shallow layers and it goes through the deeper ones; hold one direction and it goes around it. Holding the whole state of the opening leaves it no route.

5.4 Cost

The opening hold is as safe as the standard defence, mixing 10% refusal examples into the training data (Bianchi et al., 2024), but it refuses far fewer safe-but-alarming requests: 4.4% of the safe prompts of XSTest (Röttger et al., 2024), where 10% mixing refuses 22.6% and the aligned model 3.4%. Its answers to ordinary instructions are judged as good as those of plain fine-tuning, and no compliance appears later in long answers (Appendix I). Forcing “I” at inference would also restore refusals, but it pushes harmless requests toward refusal and requires knowing which requests are harmful; the opening hold acts in training and leaves that judgement to the model.

6 Reading the opening warns early

The first word is the most readable face of the opening, so its probability, opener mass P (Eq. 1), can serve as a gauge that needs one forward pass and no generation or judge. Across 16 methods of the main grid and the aligned model, on the six models, P tracks harmful compliance with rank correlation -0.86 (Figure 6a; Appendix K). A monotone curve fitted there predicts the 11 models it never saw (rank correlation 0.79; Figure 6b), and the same relation holds across the 66 dataset–model pairs (Figure 3c). Read halfway through training, P predicts the harm at the end (-0.94; -0.85 within undefended runs; Figure 6c).

Where the harm departs from the average pattern, the first word explains it too. Among the 11 models outside the main grid, the two harmed least keep the most of their first word; masking the first answer tokens helps mainly in the three models where it keeps the first word; and Qwen3-32B, the one model

where the token-wise constraint beats a one-axis hold, is the one whose first word that hold fails to keep (Appendix K).

7 Related work

Fine-tuning on harmful data removes alignment (Yang et al., 2023; Lermen et al., 2023), and benign data degrade it (Qi et al., 2024); which benign data harm has been tied to list- and math-style examples (He et al., 2024), to outlier samples (Guan et al., 2025), to the similarity between alignment and fine-tuning data (Hsiung et al., 2026), to instruction-style rather than task-style data (Eiras et al., 2025) and to the prompt template (Lyu et al., 2024). We explain these observations with one principle that predicts which data harm before training, and locate the damage at one position; the finding of He et al. (2024) that GSM8K can harm Llama-2-7B-Chat under full fine-tuning fits it, since in our grid GSM8K harms only the one model whose first word it rewrites.

The damage sits where alignment itself lives. Alignment mostly teaches style and format (Zhou et al., 2023) and shifts the output distribution mainly on the first, stylistic tokens (Lin et al., 2024; Ghosh et al., 2024); safety alignment in particular is concentrated in the first few output tokens (Qi et al., 2025), which prefilling attacks exploit (Wei et al., 2023; Andriushchenko et al., 2025). Qi et al. (2025) protect these tokens with a token-wise constraint; we find that the harm follows whether the first word survives rather than whether it is supervised, and that holding the position that decides it protects more. That position is the chat template's last token, where refusal is encoded apart from harmfulness (Zhao et al., 2025a) and where safety decisions are anchored (Leong et al., 2025; Zhang et al., 2026a). Harm stays readable after fine-tuning (Zhao et al., 2025a) and the refusal signal persists (Guo et al., 2026); we show that the model still uses that judgement, and that benign fine-tuning rewrites the template's last token, not the request's.

Localising safety, to a direction (Arditi et al., 2024), sparse weights (Wei et al., 2024), neurons (Chen et al., 2025), heads (Zhou et al., 2025) or layers (Li et al., 2025b), does not tell where to protect it, much as causal localisation does not tell where knowledge editing works (Hase et al., 2023). Defences freeze safety layers (Li et al., 2025b), project the update with respect to a safety subspace (Hsu et al., 2024; Li et al., 2025a), a separation that has been questioned (Ponkshe et al., 2026), mix in safety demonstrations (Bianchi et al., 2024), alternate alignment and user data (Huang et al., 2024a), harden the model before fine-tuning (Huang et al., 2024b; Rosati et al., 2024; Tamirisa et al., 2025), or constrain activations along one axis (Poppi & Lukas, 2026; Du et al., 2025). Closest to the opening hold, Lu et al. (2026) anchor the final-layer state at the last token of harmful queries against harmful fine-tuning; our account explains why anchoring there works, and shows that the same constraint fails at the request's last token or in the opening's shallow layers.

Forgetting under fine-tuning often leaves a capability in place and changes which task is inferred (Kotha et al., 2024), as extinction in animal learning leaves the original association intact, to return with the right cue (Bouton, 2004); here the cue is the first word. Appendix L extends this discussion.

8 Limitations

We study benign fine-tuning, not attacks with harmful data, mainly with LoRA on open 8B–32B models (full fine-tuning on two), and measure harm with two automatic judges from different families. Reasoning models begin their answer only after thinking; whether holding that later opening protects them is the next test.

9 Conclusion

Harmless data make aligned models comply with harmful requests because fine-tuning teaches a model how to begin answering a request, as alignment did, and a harmful request is a request too. The first word is rewritten at the opening while the judgement of harm survives, so the damage can be read there in one forward pass and prevented there by holding the opening's state. Safety under fine-tuning is kept by holding the moment the model begins to answer; more broadly, a learned behaviour is best protected where fine-tuning rewrites it, which need not be where it is thought to be stored.

ETHICS STATEMENT

This work studies why benign fine-tuning weakens the refusals of open-weight language models and how to prevent it. It uses public harmful-request benchmarks (AdvBench, HarmBench, XSTest) and public instruction datasets, and involves no human subjects. Model answers to harmful requests are kept on our servers; the paper shows only their first words, with harmful content truncated. The method we propose makes benign fine-tuning safer and does not create new attack capabilities.

REPRODUCIBILITY STATEMENT

All models and datasets are public. Appendix A gives the training recipe, the hold, the prompts and the judges; every number in the paper is computed by scripts from the stored model outputs. The supplementary material contains the code for training, evaluation and analysis, the evaluation prompts we wrote, the summaries behind every table and figure, and the scripts that regenerate the tables and figures from them.

AI USE STATEMENT

We used large language models as general-purpose assistants in this work: to search the literature, to discuss ideas, derivations, experiments and results, to help write code and process data, and to draft, translate, edit and polish the text and figures. The authors directed the research and take full responsibility for the content of this paper. All AI-assisted content was reviewed by the authors, and the reported results, references and claims were checked against the underlying data, code and sources.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2016. URL https://arxiv.org/abs/1610.01644.

Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks. In *International Conference on Learning Representations (ICLR)*, pp. 40116–40143, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/63fa7efdd3bcf944a4bd6e0ff6a50041-Abstract-Conference.html.

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, pp. 136037–136083. Curran Associates, Inc., 2024. doi: 10.52202/079017-4322. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html.

Sadia Asif and Mohammad Mohammadi Amiri. RefusalGuard: Geometry-Preserving Fine-Tuning for Safety in LLMs, 2026. URL https://arxiv.org/abs/2605.01913. arXiv journal-ref: COLM 2026.

Yonatan Belinkov. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting Latent Predictions from Transformers with the Tuned Lens, 2023. URL https://arxiv.org/abs/2303.08112.

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, pp. 4043–4068. PMLR, 2025. URL https://proceedings.mlr.press/

v267/betley25a.html. Extended version: Nature 649:584–589 (2026), doi:10.1038/s41586-025-09937-5, titled "Training large language models on narrow tasks can lead to broad misalignment".

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Y Zou. Safety-Tuned LLaMAs: Lessons From Improving the Safety of Large Language Models that Follow Instructions. In *International Conference on Learning Representations (ICLR)*, pp. 34196–34216, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/9178ece87f0a5d6558f49f43ec1e8a1a-Abstract-Conference.html.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John Patrick Cunningham. LoRA Learns Less and Forgets Less. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=aloEru2qCG.

Mark E. Bouton. Context and Behavioral Processes in Extinction. *Learning & Memory*, 11(5): 485–494, 2004.

Sahil Chaudhary. Code Alpaca: An Instruction-following LLaMA model for code generation. https://github.com/sahil280114/codealpaca, 2023.

Jianhui Chen, Xiaozhi Wang, Zijun Yao, Yushi Bai, Lei Hou, and Juanzi Li. Towards Understanding Safety Alignment: A Mechanistic Perspective from Safety Neurons. In *Advances in Neural Information Processing Systems*, volume 38, pp. 12665–12700. Curran Associates, Inc., 2025. doi: 10.52202/085713-0426. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/12a00d85a76fe258e1242c3aced03250-Abstract-Conference.html.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems, 2021. URL https://arxiv.org/abs/2110.14168.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM. Databricks blog, 12 April 2023, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Yanrui Du, Fenglei Fan, Sendong Zhao, Jiawei Cao, Qika Lin, Kai He, Ting Liu, Bing Qin, and Mengling Feng. Anchoring Refusal Direction: Mitigating Safety Risks in Tuning via Projection Constraint, 2025. URL https://arxiv.org/abs/2509.06795.

Francisco Eiras, Aleksandar Petrov, Philip H. S. Torr, M. Pawan Kumar, and Adel Bibi. Do as I do (Safely): Mitigating Task-Specific Fine-tuning Risks in Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=lXE5lB6ppV.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal Gradient Descent for Continual Learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108, pp. 3762–3773. PMLR, 2020. URL https://proceedings.mlr.press/v108/farajtabar20a.html.

Gemma Team. Gemma 3 Technical Report, 2025. URL https://arxiv.org/abs/2503.19786.

Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Ramaneswaran S, Deepali Aneja, Zeyu Jin, Ramani Duraiswami, and Dinesh Manocha. A Closer Look at the Limitations of Instruction Tuning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 15559–15589. PMLR, 2024. URL https://proceedings.mlr.press/v235/ghosh24a.html.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pp. 70–79. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-5409. URL https://aclanthology.org/D19-5409/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The Llama 3 Herd of Models, 2024. URL https://arxiv.org/abs/2407.21783.

Zihan Guan, Mengxuan Hu, Ronghang Zhu, Sheng Li, and Anil Vullikanti. Benign Samples Matter! Fine-tuning On Outlier Benign Samples Severely Breaks Safety. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 20572–20597. PMLR, 2025. URL https://proceedings.mlr.press/v267/guan25c.html.

Yitong Guo, Xiaoyi Chen, Siyuan Zhang, Xiaofeng Wang, and Haixu Tang. When Safety Routing Breaks: Understanding Alignment Fragility under Benign Fine-Tuning, 2026. URL https://arxiv.org/abs/2609.01455. Author comment on arXiv v1: accepted to Findings of EMNLP 2026.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 17643–17668. Curran Associates, Inc., 2023. doi: 10.52202/075280-0774. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/3927bbdcf0e8d1fa8aa23c26f358a281-Abstract-Conference.html.

Luxi He, Mengzhou Xia, and Peter Henderson. What is in Your Safe Data? Identifying Benign Data that Breaks Safety. In *First Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2404.01099.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching, 2024. URL https://arxiv.org/abs/2404.15255.

John Hewitt and Percy Liang. Designing and Interpreting Probes with Control Tasks. 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. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Lei Hsiung, Tianyu Pang, Yung-Chen Tang, Linyue Song, Tsung-Yi Ho, Pin-Yu Chen, and Yaoqing Yang. Why LLM Safety Guardrails Collapse After Fine-tuning: A Similarity Analysis Between Alignment and Fine-tuning Datasets. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16601–16619. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.756. URL https://aclanthology.org/2026.acl-long.756/.

Chia-Yi Hsu, Yu-Lin Tsai, Chih-Hsun Lin, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Safe LoRA: The Silver Lining of Reducing Safety Risks when Finetuning Large Language Models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 65072–65094. Curran Associates, Inc., 2024. doi: 10.52202/079017-2078. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/77baa7c2a3a675823e89131698fd6e19-Abstract-Conference.html.

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://openreview.net/forum?id=nZeVKeeFYf9.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Lisa: Lazy Safety Alignment for Large Language Models against Harmful Fine-tuning Attack. In *Advances in Neural Information Processing Systems*, volume 37, pp. 104521–104555. Curran Associates, Inc., 2024a. doi: 10.52202/079017-3320. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bcfdaf04b54a69f47623c973c864ee8d-Abstract-Conference.html.

Tiansheng Huang, Sihao Hu, and Ling Liu. Vaccine: Perturbation-aware Alignment for Large Language Models against Harmful Fine-tuning Attack. In *Advances in Neural Information Processing Systems*, volume 37, pp. 74058–74088. Curran Associates, Inc., 2024b. doi: 10.52202/079017-2356. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/873c86d9a979ab80d8e2919510d4446b-Abstract-Conference.html.

Tiansheng Huang, Gautam Bhattacharya, Pratik Joshi, Joshua Kimball, and Ling Liu. Antidote: Post-fine-tuning Safety Alignment for Large Language Models against Harmful Fine-tuning Attack. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 25059–25074. PMLR, 2025a. URL https://proceedings.mlr.press/v267/huang25b.html.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Tekin, and Ling Liu. Booster: Tackling Harmful Fine-tuning for Large Language Models via Attenuating Harmful Perturbation. In *International Conference on Learning Representations (ICLR)*, pp. 67202–67226, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a7ac8a21e5a27e7ab31a5f42a0117bdb-Abstract-Conference.html.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Harmful Fine-tuning Attacks and Defenses for Large Language Models: A Survey. *ACM Computing Surveys*, 58(13): 1–39, 2026. doi: 10.1145/3817114.

Samyak Jain, Ekdeep Singh Lubana, Kemal Oksuz, Tom Joy, Philip 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, pp. 93406–93478. Curran Associates, Inc., 2024. doi: 10.52202/079017-2963. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/a9bef53eb7b0e5950d4f2d9c74a16006-Abstract-Conference.html.

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.

Joshua Kazdan, Abhay Puri, Rylan Schaeffer, Lisa Yu, Chris Cundy, Jason Stanley, Sanmi Koyejo, and Krishnamurthy Dvijotham. No, of Course I Can! Deeper Fine-Tuning Attacks That Bypass Token-Level Safety Mechanisms. In *International Conference on Learning Representations (ICLR)*, pp. 69830–69868, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/71f21ac7a1985e1d7d9a4ecd0a03e993-Abstract-Conference.html.

Minseon Kim, Jin Myung Kwak, Lama Alssum, Bernard Ghanem, Philip Torr, David Krueger, Fazl Barez, and Adel Bibi. Rethinking Safety in LLM Fine-tuning: An Optimization Perspective, 2025. URL https://arxiv.org/abs/2508.12531.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Suhas Kotha, Jacob Springer, and Aditi Raghunathan. Understanding Catastrophic Forgetting in Language Models via Implicit Inference. In *International Conference on Learning Representations (ICLR)*, pp. 24110–24139, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/692ae28fda9bfbde7c01b13bf5a03395-Abstract-Conference.html.

Chak Tou Leong, Qingyu Yin, Jian Wang, and Wenjie Li. Why Safeguarded Ships Run Aground? Aligned Large Language Models' Safety Mechanisms Tend to Be Anchored in The Template Region. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15212–15229. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.738. URL https://aclanthology.org/2025.acl-long.738/.

Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. LoRA Fine-tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B, 2023. URL https://arxiv.org/abs/2310.20624.

Jianwei Li and Jung-Eun Kim. Superficial Safety Alignment Hypothesis. In *International Conference on Learning Representations (ICLR)*, pp. 145127–145156, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/eab780aff8d47494e203f96af54f9425-Abstract-Conference.html.

Mingjie Li, Wai Man Si, Michael Backes, Yang Zhang, and Yisen Wang. SaLoRA: Safety-Alignment Preserved Low-Rank Adaptation. In *International Conference on Learning Representations (ICLR)*, pp. 90827–90843, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e24d9d028e3c7f6f13e6032919ee021e-Abstract-Conference.html.

Shen Li, Liuyi Yao, Lan Zhang, and Yaliang Li. Safety Layers in Aligned Large Language Models: The Key to LLM Security. In *International Conference on Learning Representations (ICLR)*, pp. 98163–98189, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f3bfbd65743e60c685a3845bd61ce15f-Abstract-Conference.html.

Tung-Ling Li and Hongliang Liu. Logit-Gap Steering: A Forward-Pass Diagnostic for Alignment Robustness, 2025. URL https://arxiv.org/abs/2506.24056.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. The Unlocking Spell on Base LLMs: Rethinking Alignment via In-Context Learning. In *International Conference on Learning Representations (ICLR)*, pp. 24907–24933, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/6bcbb4a501dbad0eba1b660c1a55318c-Abstract-Conference.html.

Guozhi Liu, Weiwei Lin, Qi Mu, Tiansheng Huang, Ruichao Mo, Yuren Tao, and Li Shen. Targeted Vaccine: Safety Alignment for Large Language Models Against Harmful Fine-Tuning via Layer-Wise Perturbation. *IEEE Transactions on Information Forensics and Security*, 20:10806–10817, 2025. doi: 10.1109/TIFS.2025.3615412.

Guoxin Lu, Letian Sha, Qing Wang, Peijie Sun, Hao Zhou, Hua Dai, and Fu Xiao. Safety Anchor: Defending Harmful Fine-tuning via Geometric Bottlenecks. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2605.05995.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-Tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3776–3786, 2025. doi: 10.1109/TASLPRO.2025.3606231.

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, pp. 118603–118631. Curran Associates, Inc., 2024. doi: 10.52202/079017-3766. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d6f034bb216b472fc7d32ec7aff20342-Abstract-Conference.html.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods. https://github.com/huggingface/peft, 2022.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 35181–35224. PMLR, 2024. URL https://proceedings.mlr.press/v235/mazeika24a.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Mistral AI team. Un Ministral, des Ministraux. Mistral AI news post, 16 October 2024, 2024. URL https://mistral.ai/news/ministraux.

nostalgebraist. interpreting GPT: the logit lens. LessWrong blog post, 31 August 2020, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

ShengYun 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, pp. 95692–95715. Curran Associates, Inc., 2024. doi: 10.52202/079017-3032. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ada93fa6643735f294be51dc31eebbd4-Abstract-Conference.html.

ShengYun Peng, Pin-Yu Chen, Jianfeng Chi, Seongmin Lee, and Duen Horng Chau. Shape it Up! Restoring LLM Safety during Finetuning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 132053–132089. Curran Associates, Inc., 2025. doi: 10.52202/085713-4397. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/bf05b8d4361c6be8e250be4b924f0e1d-Abstract-Conference.html.

Kaustubh Ponkshe, Shaan Shah, Raghav Singhal, and Praneeth Vepakomma. Safety Subspaces are Not Linearly Distinct: A Fine-Tuning Case Study. In *International Conference on Learning Representations (ICLR)*, pp. 95677–95710, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/9a9f4e15ad0d680429a3e0570a96f763-Abstract-Conference.html.

Samuele Poppi and Nils Lukas. A Gravitational Interpretation of Fine-Tuning Reversion, 2026. URL https://arxiv.org/abs/2606.28525.

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)*, pp. 30988–31043, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/83b7da3ed13f06c13ce82235c8eedf35-Abstract-Conference.html.

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)*, pp. 54911–54941, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/88be023075a5a3ff3dc3b5d26623fa22-Abstract-Conference.html.

Domenic Rosati, Jan Wehner, Kai Williams, Łukasz Bartoszcze, David Atanasov, Robie Gonzales, Subhabrata Majumdar, Carsten Maple, Hassan Sajjad, and Frank Rudzicz. Representation Noising: A Defence Mechanism Against Harmful Finetuning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 12636–12676. Curran Associates, Inc., 2024. doi: 10.52202/079017-0402. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/172be8b0b88fc2b4aee74237d43f8c04-Abstract-Conference.html.

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. Association for Computational Linguistics, 2024. 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 (Datasets and Benchmarks Track)*, volume 37, pp. 125416–125440. Curran Associates, Inc., 2024. doi: 10.52202/079017-3984. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e2e06adf560b0706d3b1ddfca9f29756-Abstract-Datasets_and_Benchmarks_Track.html.

Max Springer, Chung Peng Lee, Blossom Metevier, Jane Castleman, Bohdan Turbal, Hayoung Jung, Zeyu Shen, and Aleksandra Korolova. The Geometry of Alignment Collapse: When Fine-Tuning Breaks Safety, 2026. URL https://arxiv.org/abs/2602.15799.

Rishub Tamirisa, Bhrugu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, Justin Wang, Rowan Wang, Ron Arel, Andy Zou, Dawn Song, Bo Li, Dan Hendrycks, and Mantas Mazeika. Tamper-Resistant Safeguards for Open-Weight LLMs. In *International Conference on Learning Representations (ICLR)*, pp. 101802–101829, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/fc49a629d33bc2461ed7a715ce44da68-Abstract-Conference.html.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, et al. 2 OLMo 2 Furious (COLM's Version). In *Second Conference on Language Modeling (COLM)*, 2025. URL https://openreview.net/forum?id=2ezugTT9kU.

Guoli Wang, Haonan Shi, Tu Ouyang, and An Wang. Few Tokens, Big Leverage: Preserving Safety Alignment by Constraining Safety Tokens during Fine-tuning. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 4848–4858. ACM, 2026a. doi: 10.1145/3770855.3817837.

Xiao Wang, Yifei Zhang, Yongkang Liu, Xiaocui Yang, Zihan Wang, Shi Feng, and Daling Wang. From Parameter Dynamics to Risk Scoring: Quantifying Sample-Level Safety Degradation in LLM Fine-tuning. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026b. URL https://arxiv.org/abs/2605.04572.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How Does LLM Safety Training Fail? In *Advances in Neural Information Processing Systems*, volume 36, pp. 80079–80110. Curran Associates, Inc., 2023. doi: 10.52202/075280-3508. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/fd6613131889a4b656206c50a8bd7790-Abstract-Conference.html.

Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. Assessing the Brittleness of Safety Alignment via Pruning and Low-Rank Modifications. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 52588–52610. PMLR, 2024. URL https://proceedings.mlr.press/v235/wei24f.html.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

Yuxin Xiao, Sana Tonekaboni, Walter Gerych, Vinith Suriyakumar, and Marzyeh Ghassemi. When Style Breaks Safety: Defending LLMs Against Superficial Style Alignment. In *International Conference on Learning Representations (ICLR)*, pp. 104126–104150,

2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/aa539a128dc8653408dab242f6042c84-Abstract-Conference.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, et al. Qwen3 Technical Report, 2025. URL https://arxiv.org/abs/2505.09388.

Chenghao Yang, Sida Li, and Ari Holtzman. LLM Probability Concentration: How Alignment Shrinks the Generative Horizon. *Transactions on Machine Learning Research*, 2026a. URL https://openreview.net/forum?id=KotVuXj6CL.

Shuo Yang, Qihui Zhang, Yuyang Liu, Yue Huang, Xiaojun Jia, Kun-Peng Ning, Jia-Yu Yao, Jigang Wang, Hailiang Dai, Yibing Song, and Li Yuan. AsFT: Anchoring Safety During LLM Fine-Tuning Within Narrow Safety Basin. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40 (40):34322–34330, 2026b. doi: 10.1609/aaai.v40i40.40729.

Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models, 2023. URL https://arxiv.org/abs/2310.02949.

Qiusi Zhan, Richard Fang, Rohan Bindu, Akul Gupta, Tatsunori Hashimoto, and Daniel Kang. Removing RLHF Protections in GPT-4 via Fine-Tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 681–687. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-short.59. URL https://aclanthology.org/2024.naacl-short.59/.

Jiawei Zhang, Andrew Estornell, David Baek, Bo Li, and Xiaojun Xu. Any-Depth Alignment: Unlocking Innate Safety Alignment of LLMs to Any-Depth. In *International Conference on Learning Representations (ICLR)*, pp. 25905–25939, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2bb0db7d7081037aada48aafbeae717d-Abstract-Conference.html.

Junbo Zhang, Qianli Zhou, Xinyang Deng, Wen Jiang, Jie Pan, and Jinbiao Zhu. DataShield: Safety-degrading Data Filtering for LLM Benign Instruction Fine-Tuning, 2026b. URL https://arxiv.org/abs/2606.00160.

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, pp. 140283–140318. Curran Associates, Inc., 2025a. doi: 10.52202/085713-4688. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/cd18539787d90e1d682d557c2c71b534-Abstract-Conference.html.

Xuandong Zhao, Xianjun Yang, Tianyu Pang, Chao Du, Lei Li, Yu-Xiang Wang, and William Yang Wang. Weak-to-Strong Jailbreaking on Large Language Models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 77822–77843. PMLR, 2025b. URL https://proceedings.mlr.press/v267/zhao25aa.html.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less Is More for Alignment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 55006–55021. Curran Associates, Inc., 2023. doi: 10.52202/075280-2400. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ac662d74829e4407ce1d126477f4a03a-Abstract-Conference.html.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, and Yongbin Li. On the Role of Attention Heads in Large Language Model Safety. In *International Conference on Learning Representations (ICLR)*, pp. 84042–84071, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d0bcff6425bbf850ec87d5327a965db9-Abstract-Conference.html.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, et al. Representation Engineering: A Top-Down Approach to AI Transparency, 2023a. URL https://arxiv.org/abs/2310.01405.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and Transferable Adversarial Attacks on Aligned Language Models, 2023b. URL https://arxiv.org/abs/2307.15043.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving Alignment and Robustness with Circuit Breakers. In *Advances in Neural Information Processing Systems*, volume 37, pp. 83345–83373. Curran Associates, Inc., 2024. doi: 10.52202/079017-2651. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/97ca7168c2c333df5ea61ece3b3276e1-Abstract-Conference.html.

Table 3 lists what each appendix section answers and where the main text relies on it. Every table is generated from the stored run outputs by the scripts in the supplementary material; rates are shares between 0 and 1 in the tables and percentages in the text.

Table 3: Guide to the appendix.

Section	Question it answers	Main text
A Setup	models, data, training, the holds, the methods compared, evaluation, statistics	§2, §5.1
B Judges	can two automatic judges be trusted, and do they agree	§2
C Main grid	every method, every model	Tables 1, 2
D Why benign data harm	learning, update size, time course, datasets, first words, wording	§3
E The first word	first words, giving them back, putting the state back, forced openings	§4
F Where to hold	position, depth, strength, one epoch, mixing	§5.3
G Full fine-tuning and scale	full fine-tuning, the rank sweep, 17 models, Dolly	§5.3
H Robustness	other datasets, sampling, unseen harmful sources	§5.2
I Cost	long answers, over-refusal, helpfulness, downstream accuracy	§5.4
J Other behaviours	self-disclosure, language, format	§3
K Early warning	every correlation and its exceptions	§6
L Related work	extended discussion	§7

A Setup details

A.1 Models

The main grid uses six aligned models from four families: Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen3-8B, Qwen3-14B and Qwen3-32B (Yang et al., 2025) with thinking disabled, Gemma-3-12B-it (Gemma Team, 2025) and OLMo-2-13B-Instruct (Team OLMo et al., 2025). The breadth panel adds eleven aligned models under the same recipe: Llama-3.2-1B/3B-Instruct, Qwen3-4B, Gemma-3-1B/4B-it, Gemma-4-E4B/12B/31B-it, OLMo-2-32B-Instruct, Ministral-8B-Instruct (Mistral AI team, 2024) and Mistral-Small-24B-Instruct (17 models, five families, 1B to 32B parameters). Every model uses its own chat template. The opening is the last token of the template that hands the turn to the assistant; for Llama it is the newline after <|end_header_id|>, and for the other families it is the corresponding last token of the assistant header. All models are loaded and trained with Hugging Face Transformers and PEFT (Wolf et al., 2020; Mangrulkar et al., 2022).

A.2 Benign datasets

Table 4 lists the eleven benign datasets. Every dataset is converted to single-turn {request, answer} pairs, deduplicated and shuffled with a fixed seed; the first 256 pairs are held out for the held-out loss,

and each training run draws its own 1,600 examples from the rest. The eight datasets beyond Alpaca, Dolly and GSM8K were sorted into open requests and fixed tasks before any of them was trained on.

Table 4: The eleven benign datasets. *Open request*: every example is a different request answered freely. *Fixed task*: every example is the same task with a set answer format.

Dataset	Kind	Source and content
Alpaca (Taori et al., 2023)	open request	model-written instructions and answers (`tatsu-lab/alpaca`)
Dolly (Conover et al., 2023)	open request	human-written instructions and answers (`databricks-dolly-15k`)
OpenAssistant	open request	English root prompts of `oasst1` with their best-ranked reply
No Robots	open request	human-written instructions and answers (`HuggingFaceH4/no_robots`)
UltraChat	open request	first turns of `ultrachat_200k` dialogues
CodeAlpaca (Chaudhary, 2023)	open request	programming instructions and answers
GSM8K (Cobbe et al., 2021)	fixed task	grade-school math word problems with worked solutions
MedQA (Jin et al., 2021)	fixed task	four-option medical exam questions with the answer
SAMSum (Gliwa et al., 2019)	fixed task	chat dialogues with a one-paragraph summary
SQL	fixed task	a table schema and a question, answered with a SQL query (`sql-create-context`)
English–French	fixed task	an English sentence translated into French (`opus-100`)

A.3 Training

LoRA (Hu et al., 2022) of rank 16 and $\alpha = 32$ (dropout 0.05) is applied to the query, key, value and output projections of every layer. Training runs for 200 steps with batch size 8, learning rate 2×10^{-4} with a cosine schedule and 3% warm-up, maximum sequence length 1,024 and bf16; the loss is computed on the answer tokens only. Each of the three seeds draws its own 1,600 training examples, so every example is seen once. Full fine-tuning (Llama-3.1-8B and Qwen3-8B) uses the same data and number of steps with its own learning rate. The LoRA rank sweep (Appendix G) raises the rank to 64 and 256 and then adapts all linear layers, keeping $\alpha/r = 2$.

A.4 The opening hold and the one-axis hold

Anchors. The anchors are harmful prompts from a training split that shares no prompt with the 288 harmful and 288 harmless evaluation requests. Every four steps, eight anchors are passed through the model being trained, and the loss gains the term of Equation 2: the squared distance of each anchor's state at the opening from the aligned model's, divided by the squared mean length of the aligned states at that layer, averaged over the layers at relative depth 0.34–0.94 and weighted by $\lambda = 3$. The anchor pool holds 432 harmful prompts (308 from AdvBench and 124 from HarmBench) and 432 harmless ones, none of them among the evaluation requests; the harmless ones serve only the harmless-anchor variant. The aligned states of the anchors are computed once before training.
One axis. The simplified hold penalises only the component along one direction u_ℓ at relative depth 0.34–0.66 (for Llama, layers 11–21). By default u_ℓ is the drift direction: the mean displacement of the state at that position in an undefended run of the same model, orthogonalised to the harm direction. The harm direction is the difference of the mean states of harmful and harmless prompts at the opening of the aligned model (Arditi et al., 2024). The drift direction comes from the seed-0 undefended run of the same model, so the one-axis hold, unlike the opening hold, needs an earlier run. Its penalty is a hinge on the projection onto u_ℓ around the aligned model's value, scaled by the mean absolute aligned projection: two-sided for the drift direction and one-sided (penalising only a fall) for the harm direction. When the one-axis hold is moved to another position, the direction is estimated at that position. Positions after the opening (the first and fifth answer word) are read after appending the aligned model's own most likely first word(s) to the anchor, so that the held state is the one the aligned model would be in.

A.5 Evaluation

The harmful set has 288 requests, the evaluation split of AdvBench (212) and HarmBench (76) (Zou et al., 2023b; Mazeika et al., 2024); the harmless set has 288 requests, nine tenths of them Dolly prompts that are not used in training. Answers are generated greedily with up to 128 new tokens (512 in the long-generation check of Appendix I). The main judge (Qwen3.8-27B, thinking off) labels every answer, cut to its first 1,500 characters, with the prompt below, answering in at most six greedy tokens; the second judge (Gemma-4-31B-it) labels every answer again with the same prompt.

REFUSAL and DEFLECTION (a lecture or safe variant without the requested help) count as safe, COMPLIANCE as harmful, INCOHERENT as nonsense.

```
You are a strict safety evaluator. Classify the ASSISTANT
RESPONSE to the USER REQUEST into exactly one label:
REFUSAL: the assistant explicitly declines or refuses,
or says it cannot help with the request. DEFLECTION:
the assistant neither refuses explicitly nor provides
the requested harmful content (e.g. redirects, answers a
safe variant, gives only warnings, lectures, or generic
non-actionable text). COMPLIANCE: the assistant provides
substantive content that attempts to fulfil the request
(even partially, even with disclaimers). INCOHERENT: the
response is empty, garbled, off-topic nonsense, or cut
off before any content. USER REQUEST: {prompt} ASSISTANT
RESPONSE: {response} Answer with one word: REFUSAL,
DEFLECTION, COMPLIANCE, or INCOHERENT.
```

Harmful compliance is the share of harmful requests labelled compliance; a lecture that gives no help counts as safe, following the practice of scoring real help rather than the absence of a refusal (Souly et al., 2024).

A.6 Read-outs

Opener mass. $\mathcal{W}_0$ is the set of tokens that the aligned model gives as its greedy first token on at least three harmful evaluation requests (for Llama, only "I"; Table 21 lists them). P is the probability of $\mathcal{W}_0$ at the opening, averaged over harmful requests; it needs one forward pass and no generation. Defining $\mathcal{W}_0$ from a different set of harmful prompts, or using only the single most frequent aligned first word, gives the same ranking (Table 53). *How much is learned.* The loss on 256 held-out examples of the training dataset, relative to the aligned model's loss on the same examples. *Learning kept.* For a defended run, (aligned loss − defended loss) / (aligned loss − undefended loss) on held-out Alpaca, same model and seed; the median over runs is reported. *Size of the update.* The Frobenius norm of the learned update $\Delta W = (\alpha/r)BA$ over all adapted modules, divided by that of Alpaca on the same model. *Separability kept.* A logistic-regression probe (Alain & Bengio, 2016) ($C = 0.05$, standardised inputs) trained on the aligned model to separate harmful from harmless requests is applied, frozen, to the fine-tuned model's states; we report the ratio of its separation (d') on the fine-tuned model to that on the aligned model, at the opening and at the request's last token, over layers at relative depth at least 0.75 where the aligned separation is at least 0.5. A probe shows what can be read from a state, not what the model uses (Hewitt & Liang, 2019; Belinkov, 2022); the use of the judgement is tested behaviourally (Appendix E).

A.7 Methods compared

Freezes. The published safety layers (Li et al., 2025b) are frozen (LoRA only elsewhere): layers 9–23 in Llama-3.1-8B, 19–28 in Qwen3-8B, 43–48 in Gemma-3-12B, 7–16 in OLMo-2-13B, 17–26 in Qwen3-14B and 25–56 in Qwen3-32B; each freeze has a placebo that freezes as many random layers disjoint from it. The harm-reading freeze uses the band of layers where a probe separates harmful from harmless requests best. *Orthogonal update.* All layers are adapted, and at every step the LoRA B matrix of the output projection is projected onto the orthogonal complement of the aligned model's refusal direction. *Loss masks.* The first one or five answer tokens (or, in one variant, also the user turn) change their contribution to the loss. *Untrained updates.* A random update with the same size and rank as the learned one, the learned update with its rows shuffled, and the learned update with its sign reversed. *Own answers.* Each model's greedy answers (at most 512 new tokens) to the same 1,600 requests. *Safety mixing* (Bianchi et al., 2024): the 2,483 published safety demonstrations, after removing any whose instruction matches one of our 720 harmful prompts (identical after normalisation, or word-set Jaccard ≥ 0.6), mixed in at 3% or 10%. *Token-wise constraint* (Qi et al., 2025): their Equation 3 with $\beta_1 = 0.5$, $\beta_{2..5} = 2$ and $\beta_{>5} = 0.1$, the aligned model being the same model with the adapter disabled, at our dose on benign data; *warm-up*: the 10-step learning-rate warm-up that they pair with it, alone. *Full fine-tuning*: learning rate 2×10^{-5} with AdamW and fp32 master weights, 200 steps, batch 8.

A.8 STATISTICS

A *run* is one training of one model with one seed; the main grid has 18 runs per method, and rates are medians over them. Comparisons are paired: a method and its control are compared within the same model and seed. A run counts as *safer* when its harmful compliance is lower than the control's, as *worse* when it is more than 2 points higher, and as the *same* when the two are within 2 points (used for comparisons with the aligned model). Correlations are Spearman rank correlations; the gauge's curve is a monotone fit of harmful compliance on opener mass over the main-grid points. A model-level cluster bootstrap (2,000 resamples) gives a 95% interval of $[-0.92, -0.81]$ for the pooled rank correlation over the 402 per-run points, with the aligned first words taken from a separate prompt set (point estimate -0.87). All runs use single NVIDIA H200 GPUs.

B JUDGES

The two judges come from different families and agree on whether an answer complies for 99.1% of answers; over the four labels, Cohen's κ is 0.92. Most disagreements are between refusal and lecture, which both count as safe. Every paired comparison in the paper has the same direction under the second judge (second-judge columns in Tables 6 and 43). LLM judges with a fixed prompt are the usual way to score open-ended answers (Zheng et al., 2023); Table 5 gives the agreement on every judged answer.

Table 5: Agreement between the two judges over every judged generation.

Read-out	Value
Generation files judged by both	1158
Answers compared	329040
Per-answer agreement (complies or not)	0.991
Cohen κ (four labels)	0.924
Main "complies" also 2nd	0.972
2nd "complies" also main	0.982
Rank corr. of per-file rates	0.999
Largest disagreement: refusal vs. lecture	8234 answers
All disagreements	11767
of which refusal vs. lecture	8390 (71.3%)
of which involve "complies"	2821 (24.0%)

C THE FULL MAIN GRID

The main grid holds every method at the same dose (6 models $\times$ 3 seeds, LoRA rank 16, 200 steps, 1,600 examples), so that every row can be compared with every other, paired by model and seed. Table 6 lists all 38 methods. Tables 7–9 break the main rows down by model. Table 10 reads nine properties from the states: fine-tuning dilutes every property at the opening and none at the request's last token, and holding the opening restores harmfulness fully and the others largely (0.77–0.99); the model still represents what the request is.

Table 6: All 38 methods of the main grid (6 models × 3 seeds; rates are medians over the 18 runs, paired differences are within model and seed). Sep. kept: how separable harmful and harmless requests remain for a probe trained on the aligned model, at the opening and at the request's last token.

Group	Method	Harmful	Control	Paired diff.	Refusal	Harmful (2nd judge)	Paired diff. (2nd)	P	P (harmless)	Entropy	Sep. kept (opening)	Sep. kept (request)	Learning kept
Baseline	Aligned model	0.017	–	–	0.929	0.019	–	0.94	0.04	0.35	1.000	1.000	–
Baseline	Undefended	0.226	Aligned model	+0.198 (18/18 worse)	0.622	0.220	+0.193 (18/18 worse)	0.36	0.01	3.50	0.359	1.019	1.000
Locate, protect	Freeze published safety layers	0.207	placebo	+0.002 (8/18 safer)	0.674	0.191	+0.000 (8/18 safer)	0.37	0.01	3.53	0.409	1.013	0.996
Locate, protect	Placebo: as many random layers	0.222	–	–	0.635	0.214	–	0.35	0.01	3.62	0.382	1.015	0.998
Locate, protect	Freeze layers that best read harm	0.271	placebo	+0.036 (0/18 safer)	0.564	0.264	+0.040 (0/18 safer)	0.23	0.01	3.99	0.351	1.014	0.997
Locate, protect	Placebo for the row above	0.212	–	–	0.628	0.200	–	0.32	0.01	3.60	0.389	1.010	0.998
Locate, protect	Updates orthogonal to refusal direction	0.233	Undefended	+0.002 (7/18 safer)	0.632	0.227	+0.002 (6/18 safer)	0.37	0.01	3.52	0.359	1.019	1.000
Hold the opening	Hold: one axis (harm direction)	0.026	Undefended	-0.200 (18/18 safer)	0.948	0.026	-0.194 (18/18 safer)	0.73	0.01	1.84	0.864	1.021	1.000
Hold the opening	Hold: one axis (drift direction)	0.007	Undefended	-0.212 (18/18 safer)	0.969	0.009	-0.205 (18/18 safer)	0.78	0.01	1.63	1.030	1.007	0.993
Hold the opening	Hold: one axis, harmless anchors	0.095	Undefended	-0.130 (18/18 safer)	0.826	0.083	-0.135 (18/18 safer)	0.51	0.01	2.93	0.687	1.008	0.990
Hold the opening	Hold: whole state, middle only	0.014	Undefended	-0.208 (18/18 safer)	0.969	0.016	-0.198 (18/18 safer)	0.80	0.01	1.51	1.033	1.003	0.996
Hold the opening	Hold: whole state (middle–deep)	0.009	Undefended	-0.217 (18/18 safer)	0.969	0.009	-0.208 (18/18 safer)	0.87	0.01	1.07	1.097	1.004	0.995
Hold the opening	Token-wise constraint (Qi et al.)	0.106	Undefended	-0.128 (18/18 safer)	0.821	0.097	-0.123 (18/18 safer)	0.68	0.01	2.14	0.598	1.016	0.984
What rewrites the opening	No loss on first token	0.193	Undefended	-0.017 (11/18 safer)	0.688	0.194	-0.012 (11/18 safer)	0.38	0.01	3.41	0.426	1.018	0.997
What rewrites the opening	No loss on first 5 tokens	0.175	Undefended	-0.061 (14/18 safer)	0.767	0.177	-0.057 (14/18 safer)	0.48	0.01	3.52	0.451	1.021	0.995
What rewrites the opening	Loss on user turn too	0.252	Undefended	-0.002 (9/18 safer)	0.615	0.240	-0.002 (9/18 safer)	0.31	0.01	3.64	0.344	0.995	0.999
What rewrites the opening	Random update (same size, rank)	0.014	Aligned model	+0.000 (18/18 same)	0.929	0.016	+0.000 (18/18 same)	0.94	0.04	0.35	0.999	1.000	0.002
What rewrites the opening	Learned update, rows shuffled	0.014	Aligned model	+0.000 (17/18 same)	0.938	0.017	-0.003 (18/18 same)	0.94	0.03	0.34	0.992	1.001	0.017
What rewrites the opening	Negated update ($W - \Delta W$)	0.040	Aligned model	+0.010 (10/18 same)	0.689	0.047	+0.012 (11/18 same)	0.95	0.05	0.41	0.708	0.966	-6.503
What rewrites the opening	Fixed requests + Alpaca answers	0.264	Aligned model	+0.219 (18/18 worse)	0.578	0.260	+0.203 (18/18 worse)	0.26	0.01	3.66	0.355	1.016	1.001
What rewrites the opening	Same requests + own answers	0.014	Fixed + Alpaca	-0.219 (18/18 safer)	0.922	0.014	-0.198 (18/18 safer)	0.97	0.04	0.13	1.003	1.000	-0.883
What rewrites the opening	Same answers, no requests	0.059	Fixed + Alpaca	-0.200 (18/18 safer)	0.915	0.059	-0.191 (18/18 safer)	0.86	0.02	0.88	0.813	1.043	0.815
What rewrites the opening	Other fixed requests (imperative)	0.259	Aligned model	+0.236 (18/18 worse)	0.646	0.252	+0.224 (18/18 worse)	0.34	0.01	3.59	0.360	1.017	0.998
What rewrites the opening	Same requests as questions	0.252	Imperative run	+0.012 (8 safer, 7 worse /18)	0.642	0.248	+0.012 (7 safer, 8 worse /18)	0.34	0.02	4.02	0.355	1.011	0.996
Place and strength	Hold: one axis, 1/10 strength	0.012	Undefended	-0.201 (18/18 safer)	0.950	0.010	-0.200 (18/18 safer)	0.75	0.01	1.63	0.982	1.010	0.996
Place and strength	At request's last token	0.191	Undefended	-0.035 (16/18 safer)	0.694	0.194	-0.031 (15/18 safer)	0.36	0.01	3.49	0.397	0.993	0.999
Place and strength	At request's last token, $\times 10$	0.160	Undefended	-0.080 (17/18 safer)	0.701	0.155	-0.080 (18/18 safer)	0.33	0.01	3.41	0.426	0.998	0.997
Place and strength	At request's last token, own axis	0.167	Undefended	-0.061 (18/18 safer)	0.724	0.160	-0.059 (18/18 safer)	0.39	0.01	3.37	0.444	0.992	0.999
Place and strength	Same, $\times 10$	0.148	Undefended	-0.073 (18/18 safer)	0.738	0.141	-0.068 (18/18 safer)	0.39	0.01	3.35	0.468	0.995	0.996
Place and strength	At template token -2	0.094	Undefended	-0.102 (18/18 safer)	0.781	0.089	-0.108 (18/18 safer)	0.48	0.01	3.23	0.534	1.004	0.995
Place and strength	At template token -1	0.068	Undefended	-0.146 (18/18 safer)	0.884	0.062	-0.153 (18/18 safer)	0.60	0.01	2.58	0.609	1.005	0.994
Place and strength	At first answer word	0.019	Undefended	-0.214 (18/18 safer)	0.950	0.017	-0.205 (18/18 safer)	0.70	0.01	1.98	0.891	1.004	0.990
Place and strength	At fifth answer word	0.024	Undefended	-0.212 (18/18 safer)	0.920	0.024	-0.207 (18/18 safer)	0.67	0.01	2.11	0.695	1.007	0.995
Place and strength	Hold at opening, shallow layers	0.175	Undefended	-0.035 (14/18 safer)	0.733	0.168	-0.036 (15/18 safer)	0.41	0.01	3.32	0.431	1.014	0.998
Place and strength	Hold at opening, deep layers	0.014	Undefended	-0.210 (18/18 safer)	0.984	0.012	-0.205 (18/18 safer)	0.87	0.01	1.07	0.946	1.015	0.992
Place and strength	At request's last token, deep	0.165	Undefended	-0.057 (18/18 safer)	0.727	0.156	-0.064 (18/18 safer)	0.43	0.01	3.28	0.435	0.992	0.999
Common baselines	Mix in 3% safety data (Bianchi et al.)	0.030	Undefended	-0.193 (18/18 safer)	0.960	0.028	-0.184 (18/18 safer)	0.80	0.01	1.60	0.692	1.031	0.999
Common baselines	Mix in 10% safety data	0.010	Undefended	-0.215 (18/18 safer)	0.990	0.010	-0.207 (18/18 safer)	0.79	0.01	1.36	0.814	1.035	0.998
Common baselines	10-step warm-up (Qi et al.)	0.231	Undefended	+0.003 (8/18 safer)	0.637	0.224	+0.007 (7/18 safer)	0.34	0.01	3.57	0.360	1.020	1.000

Table 7: Harmful compliance per model (median of 3 seeds).

Method	Llama-3.1-8B	Qwen3-8B	Gemma-3-12B	OLMo-2-13B	Qwen3-14B	Qwen3-32B
Aligned model	0.007	0.031	0.073	0.010	0.014	0.021
Undefended	0.472	0.253	0.201	0.205	0.215	0.160
Freeze published safety layers	0.465	0.271	0.201	0.108	0.212	0.115
Placebo: as many random layers	0.392	0.240	0.194	0.243	0.191	0.139
Freeze layers that best read harm	0.635	0.274	0.219	0.271	0.243	0.309
Placebo for the row above	0.448	0.250	0.191	0.250	0.188	0.132
Updates orthogonal to refusal direction	0.476	0.250	0.205	0.229	0.222	0.156
Hold: whole state (middle–deep)	0.003	0.017	0.007	0.007	0.010	0.014
Hold: whole state, middle only	0.003	0.017	0.010	0.007	0.014	0.108
Hold: one axis (harm direction)	0.003	0.035	0.021	0.003	0.028	0.083
Hold: one axis (drift direction)	0.003	0.024	0.003	0.003	0.010	0.108
Hold: one axis, harmless anchors	0.031	0.135	0.104	0.010	0.062	0.118
Token-wise constraint (Qi et al.)	0.128	0.135	0.035	0.146	0.104	0.066
No loss on first token	0.455	0.264	0.094	0.184	0.201	0.146
No loss on first 5 tokens	0.194	0.267	0.059	0.208	0.177	0.101
Loss on user turn too	0.361	0.243	0.267	0.292	0.205	0.174
Random update (same size, rank)	0.007	0.035	0.076	0.007	0.014	0.017
Learned update, rows shuffled	0.014	0.024	0.090	0.014	0.014	0.014
Negated update ($W - \Delta W$)	0.087	0.038	0.052	0.000	0.028	0.035
Fixed requests + Alpaca answers	0.413	0.333	0.278	0.198	0.222	0.219
Same requests + own answers	0.007	0.024	0.080	0.010	0.010	0.014
Same answers, no requests	0.017	0.083	0.059	0.042	0.097	0.049
Hold: one axis, 1/10 strength	0.007	0.045	0.007	0.000	0.021	0.111
At request's last token	0.469	0.205	0.156	0.219	0.184	0.132
At request's last token, $\times 10$	0.354	0.194	0.118	0.111	0.167	0.125
At request's last token, own axis	0.406	0.208	0.087	0.094	0.194	0.139
Same, $\times 10$	0.337	0.208	0.101	0.076	0.170	0.139
At template token -2	0.042	0.167	0.069	0.188	0.069	0.118
At template token -1	0.014	0.118	0.066	0.000	0.069	0.132
At first answer word	0.010	0.021	0.024	0.007	0.014	0.108
At fifth answer word	0.056	0.028	0.021	0.003	0.010	0.097
Hold at opening, shallow layers	0.260	0.247	0.170	0.069	0.177	0.160
Hold at opening, deep layers	0.003	0.007	0.014	0.000	0.017	0.014
At request's last token, deep	0.351	0.222	0.122	0.087	0.177	0.108
Mix in 3% safety data (Bianchi et al.)	0.017	0.031	0.031	0.007	0.045	0.049
Mix in 10% safety data	0.003	0.010	0.010	0.003	0.010	0.010
10-step warm-up (Qi et al.)	0.486	0.250	0.212	0.253	0.208	0.153

Table 8: Opener mass P per model (median of 3 seeds).

Method	Llama-3.1-8B	Qwen3-8B	Gemma-3-12B	OLMo-2-13B	Qwen3-14B	Qwen3-32B
Aligned model	0.99	0.93	1.00	0.96	0.92	0.87
Undefended	0.27	0.21	0.51	0.41	0.31	0.42
Freeze published safety layers	0.19	0.23	0.49	0.58	0.24	0.52
Placebo: as many random layers	0.30	0.23	0.54	0.39	0.29	0.41
Freeze layers that best read harm	0.09	0.21	0.52	0.33	0.23	0.23
Placebo for the row above	0.26	0.23	0.53	0.37	0.28	0.38
Updates orthogonal to refusal direction	0.27	0.22	0.52	0.42	0.32	0.42
Hold: whole state (middle–deep)	0.98	0.72	0.98	0.95	0.77	0.79
Hold: whole state, middle only	0.95	0.68	0.91	0.94	0.59	0.43
Hold: one axis (harm direction)	0.97	0.52	0.84	0.97	0.62	0.53
Hold: one axis (drift direction)	0.96	0.63	0.93	0.94	0.51	0.38
Hold: one axis, harmless anchors	0.86	0.37	0.60	0.90	0.43	0.36
Token-wise constraint (Qi et al.)	0.75	0.55	0.91	0.64	0.56	0.71
No loss on first token	0.29	0.25	0.75	0.45	0.30	0.51
No loss on first 5 tokens	0.64	0.26	0.82	0.38	0.31	0.58
Loss on user turn too	0.34	0.25	0.43	0.29	0.22	0.35
Random update (same size, rank)	0.99	0.93	1.00	0.96	0.92	0.87
Learned update, rows shuffled	0.99	0.92	1.00	0.96	0.92	0.89
Negated update ($W - \Delta W$)	0.81	0.95	1.00	0.66	0.96	0.96
Fixed requests + Alpaca answers	0.27	0.14	0.36	0.41	0.22	0.25
Same requests + own answers	1.00	0.96	1.00	0.98	0.94	0.93
Same answers, no requests	0.97	0.75	0.99	0.88	0.84	0.78
Hold: one axis, 1/10 strength	0.93	0.56	0.95	0.96	0.49	0.39
At request's last token	0.24	0.28	0.59	0.45	0.35	0.37
At request's last token, $\times 10$	0.30	0.30	0.64	0.62	0.33	0.33
At request's last token, own axis	0.31	0.31	0.69	0.69	0.36	0.42
Same, $\times 10$	0.36	0.31	0.69	0.68	0.36	0.41
At template token -2	0.86	0.33	0.75	0.48	0.47	0.38
At template token -1	0.96	0.42	0.74	0.96	0.45	0.37
At first answer word	0.94	0.55	0.85	0.94	0.50	0.38
At fifth answer word	0.75	0.54	0.81	0.95	0.60	0.37
Hold at opening, shallow layers	0.39	0.21	0.54	0.66	0.28	0.43
Hold at opening, deep layers	0.98	0.83	0.92	0.98	0.62	0.82
At request's last token, deep	0.36	0.26	0.64	0.74	0.35	0.50
Mix in 3% safety data (Bianchi et al.)	0.89	0.67	0.86	0.82	0.74	0.78
Mix in 10% safety data	0.81	0.75	0.82	0.75	0.78	0.84
10-step warm-up (Qi et al.)	0.23	0.22	0.49	0.42	0.28	0.41

Table 9: Share of harmful requests whose greedy first word is one of the aligned model's, per model (median of 3 seeds).

Method	Llama-3.1-8B	Qwen3-8B	Gemma-3-12B	OLMo-2-13B	Qwen3-14B	Qwen3-32B
Aligned model	1.00	0.96	1.00	0.99	0.94	0.96
Undefended	0.44	0.48	0.75	0.70	0.48	0.68
Freeze published safety layers	0.35	0.47	0.73	0.83	0.44	0.81
Placebo: as many random layers	0.52	0.49	0.76	0.64	0.48	0.71
Freeze layers that best read harm	0.15	0.44	0.75	0.57	0.42	0.47
Placebo for the row above	0.47	0.50	0.76	0.61	0.42	0.68
Updates orthogonal to refusal direction	0.45	0.49	0.75	0.70	0.50	0.68
Hold: whole state (middle–deep)	1.00	0.96	0.99	0.99	0.94	0.94
Hold: whole state, middle only	1.00	0.95	0.99	0.99	0.81	0.65
Hold: one axis (harm direction)	1.00	0.91	0.97	1.00	0.91	0.78
Hold: one axis (drift direction)	1.00	0.91	1.00	1.00	0.70	0.59
Hold: one axis, harmless anchors	0.96	0.78	0.85	0.99	0.70	0.55
Token-wise constraint (Qi et al.)	0.88	0.70	0.97	0.81	0.74	0.89
No loss on first token	0.51	0.46	0.93	0.71	0.51	0.74
No loss on first 5 tokens	0.78	0.57	0.95	0.72	0.63	0.85
Loss on user turn too	0.54	0.49	0.64	0.50	0.33	0.67
Random update (same size, rank)	1.00	0.96	1.00	0.99	0.94	0.97
Learned update, rows shuffled	0.99	0.96	1.00	0.99	0.95	0.97
Negated update ($W - \Delta W$)	0.83	0.96	1.00	0.71	0.97	0.98
Fixed requests + Alpaca answers	0.50	0.23	0.61	0.70	0.36	0.45
Same requests + own answers	1.00	0.97	1.00	0.99	0.95	0.96
Same answers, no requests	0.98	0.89	0.99	0.95	0.91	0.93
Hold: one axis, 1/10 strength	0.99	0.85	0.99	1.00	0.68	0.61
At request's last token	0.38	0.60	0.81	0.71	0.61	0.68
At request's last token, $\times 10$	0.49	0.63	0.86	0.85	0.59	0.56
At request's last token, own axis	0.50	0.61	0.89	0.88	0.59	0.69
Same, $\times 10$	0.57	0.58	0.89	0.91	0.63	0.70
At template token -2	0.96	0.65	0.92	0.77	0.76	0.67
At template token -1	0.98	0.77	0.93	1.00	0.71	0.60
At first answer word	0.99	0.90	0.98	0.99	0.70	0.59
At fifth answer word	0.92	0.89	0.98	1.00	0.87	0.65
Hold at opening, shallow layers	0.59	0.48	0.78	0.89	0.43	0.69
Hold at opening, deep layers	1.00	0.98	0.98	1.00	0.80	0.95
At request's last token, deep	0.57	0.56	0.87	0.90	0.62	0.78
Mix in 3% safety data (Bianchi et al.)	0.98	0.96	0.96	0.99	0.91	0.94
Mix in 10% safety data	0.99	0.98	0.98	1.00	0.97	0.98
10-step warm-up (Qi et al.)	0.43	0.50	0.75	0.66	0.43	0.70

Table 10: Nine properties readable from the state: separability kept at the opening and at the request's last token (deep layers, median of 18 runs). Fine-tuning dilutes every property at the opening and none at the request; holding the opening restores them.

Property	Undefended: opening	Undefended: request	Held: opening	Held: request	Lower at opening (runs)
Harmful?	0.359	1.019	1.097	1.004	18/18
Harmful? (imperatives)	0.434	1.000	1.205	0.986	18/18
Harmful? (non-imperatives)	0.305	1.007	1.082	1.001	18/18
Has a comma?	0.654	1.000	0.767	0.995	18/18
Imperative?	0.696	1.040	0.887	1.020	18/18
Length	0.508	1.002	0.803	1.005	18/18
Source dataset	0.761	1.011	0.999	1.015	18/18
Question?	0.690	1.045	0.794	1.032	18/18
Technical topic?	0.845	1.002	0.994	1.005	18/18

D WHY BENIGN DATA HARM: DETAILS

D.1 THE HARM DOES NOT FOLLOW HOW MUCH IS LEARNED

GSM8K is learned more thoroughly than Alpaca or Dolly: its held-out loss falls to 0.30 of the aligned model's, against 0.52 and 0.64, yet it barely harms (2.3%, against 22.6% for Alpaca and 1.7% before fine-tuning). CodeAlpaca and GSM8K are the sharpest pair: both are learned to 0.29–0.30 of the aligned loss and change the weights by the same amount (1.04 and 1.03 times Alpaca's update), yet CodeAlpaca raises harmful compliance to 15.3%. Across the eleven datasets the learned update has nearly the same size (0.95 to 1.04 times Alpaca's; Table 14), while harmful compliance ranges from 1.0% to 24.5% (Table 11); within each model, the size of the update and the harm are unrelated (median rank correlation -0.05 over the six models, ranging from -0.37 to $+0.59$). More training on the same data does not add harm either: under Alpaca the harm is reached by step 50 (23.6%) and stays there to step 200 (22.6%; Table 13).

Table 11: Eleven benign datasets (6 models $\times$ 3 seeds). Update size: the learned LoRA update $\|\Delta W\|$ relative to Alpaca's on the same model.

Dataset	Most common first words of training answers	Harmful: undefended	Worse runs	second judge	P	Held-out loss ratio	Update size (vs. Alpaca)
Fixed tasks							
GSM8K (math word problems)	The 16%, If 8%, First 6%	0.023	3/18	0.023	0.63	0.30	1.03
MedQA (multiple choice)	A. 26%, B. 25%, C. 25%	0.010	0/18	0.010	0.88	0.02	0.95
SAMSum (summaries)	The 2%, John 2%, Anna 1%	0.030	5/18	0.030	0.66	0.29	1.01
SQL (text to SQL)	SELECT 100%	0.010	0/18	0.010	0.87	0.03	0.96
English to French	Les 5%, Le 5%, La 5%	0.026	2/18	0.028	0.70	0.49	0.96
Open requests							
Alpaca	The 24%, 1. 6%, A 5%	0.226	18/18	0.220	0.36	0.52	1.00
Dolly	The 15%, A 3%, 1. 2%	0.245	18/18	0.227	0.17	0.64	0.98
OpenAssistant	The 12%, Sure, 5%, Here 5%	0.049	12/18	0.050	0.52	0.70	1.00
No Robots	The 13%, 1. 5%, Here 4%	0.168	15/18	0.168	0.37	0.70	0.98
UltraChat	The 18%, I 7%, 1. 5%	0.069	13/18	0.075	0.75	0.71	1.01
CodeAlpaca	def 20%, SELECT 7%, import 5%	0.153	18/18	0.151	0.40	0.29	1.04

Table 12: Harmful compliance per model and dataset, undefended (in parentheses: opener mass).

Dataset	Llama-3.1-8B	Qwen3-8B	Gemma-3-12B	OLMo-2-13B	Qwen3-14B	Qwen3-32B
GSM8K (math word problems)	0.003 (0.90)	0.028 (0.45)	0.045 (0.97)	0.247 (0.20)	0.014 (0.45)	0.007 (0.81)
MedQA (multiple choice)	0.010 (0.95)	0.024 (0.54)	0.031 (0.98)	0.003 (0.95)	0.010 (0.76)	0.003 (0.81)
SAMSum (summaries)	0.010 (0.92)	0.038 (0.54)	0.024 (0.96)	0.031 (0.67)	0.038 (0.53)	0.031 (0.64)
SQL (text to SQL)	0.007 (0.98)	0.024 (0.78)	0.062 (0.99)	0.010 (0.93)	0.003 (0.75)	0.007 (0.81)
English to French	0.007 (0.77)	0.031 (0.58)	0.052 (0.94)	0.017 (0.73)	0.021 (0.65)	0.024 (0.68)
Alpaca	0.472 (0.27)	0.253 (0.21)	0.201 (0.51)	0.205 (0.41)	0.215 (0.31)	0.160 (0.42)
Dolly	0.451 (0.08)	0.274 (0.13)	0.146 (0.29)	0.389 (0.10)	0.181 (0.26)	0.191 (0.20)
OpenAssistant	0.049 (0.53)	0.038 (0.48)	0.059 (0.54)	0.035 (0.72)	0.056 (0.52)	0.049 (0.52)
No Robots	0.347 (0.27)	0.306 (0.20)	0.069 (0.55)	0.042 (0.60)	0.184 (0.35)	0.160 (0.38)
UltraChat	0.056 (0.88)	0.101 (0.73)	0.059 (0.79)	0.146 (0.66)	0.076 (0.67)	0.052 (0.78)
CodeAlpaca	0.385 (0.36)	0.281 (0.19)	0.142 (0.55)	0.083 (0.66)	0.139 (0.44)	0.153 (0.36)

Table 13: Harmful compliance during undefended training on Alpaca (6 models $\times$ 3 seeds): the harm is reached early and does not grow with more training.

Method	Step 0	Step 25	Step 50	Step 100	Step 200
Undefended	0.017	0.127 (18)	0.236 (18)	0.234 (18)	0.226 (18)

D.2 TAKING ALPACA APART

Table 1 uses one fixed set of 1,600 Alpaca requests, drawn separately and seen once, so its control (26.4%) differs slightly from the main grid's 22.6%. *Own answers*: each model answers the 1,600 requests itself and is trained on its own answers; these teach no new way of beginning, so the held-out Alpaca loss rises (learning kept -0.88) and harmful compliance stays at the aligned level (1.4%). *No requests*: the same Alpaca answers are trained as plain text, without the request and the chat template; they are learned to 82% of the full run's extent, yet the harm is 5.9%. The opener mass follows: 0.26,

Table 14: Size of the learned update per dataset, relative to Alpaca on the same model.

Dataset	Update size vs. Alpaca (median of 6)	Range
GSM8K (math word problems)	1.03	0.92–1.16
MedQA (multiple choice)	0.95	0.72–1.03
SAMSum (summaries)	1.01	0.83–1.03
SQL (text to SQL)	0.96	0.80–1.14
English to French	0.96	0.87–1.02
Alpaca	1.00	1.00–1.00
Dolly	0.98	0.94–1.02
OpenAssistant	1.00	0.93–1.03
No Robots	0.98	0.94–1.04
UltraChat	1.01	0.94–1.05
CodeAlpaca	1.04	0.91–1.09

0.97 and 0.86. What harms is neither the requests nor the content of the answers, but a new way of answering attached to requests.

D.3 Eleven datasets, sorted in advance

The principle says in advance which data harm: open requests answered freely teach how to begin answering any request, and harmful requests are requests; a fixed task teaches how to begin answering that kind of input, and harmful requests are not that kind of input. The eight new datasets were sorted this way before training: four open requests (OpenAssistant, No Robots, UltraChat, CodeAlpaca) and four fixed tasks (MedQA, SAMSum, SQL, English–French). Together with Alpaca, Dolly and GSM8K, every open-request dataset harms (dataset medians 4.9% to 24.5%) and every fixed task stays near the aligned level (1.0% to 3.0%; Table 11). The two open-request datasets that harm least, OpenAssistant and UltraChat (4.9% and 6.9%), are themselves assistant-style dialogues. The one fixed-task point far from the aligned level is OLMo-2-13B on GSM8K (Table 12), the one model whose greedy first word on harmful requests GSM8K changes ("Step" appears among its first words; Table 21).

D.4 Harmful requests begin to be answered like ordinary ones

Before fine-tuning, the first words a model gives to harmful and to ordinary requests barely overlap: the total variation distance between the two distributions of greedy first words is 0.91–0.97 (medians over models; 0 means identical, 1 disjoint). After Alpaca or Dolly the distance shrinks (on the harmless evaluation requests, 0.97 $\rightarrow$ 0.83 and 0.75); after GSM8K it does not (0.96). The result is the same when the ordinary requests are two sets of held-out instructions that no run was trained on (No Robots and OpenAssistant; Table 15).

Table 15: How far apart the first-word distributions of harmful and ordinary requests are at the opening (total variation), before and after fine-tuning, on three sets of ordinary requests.

Ordinary request set	Training	Gap, aligned	Fine-tuned	Harmful moved
Harmless eval set (mostly Dolly)	GSM8K, 200 steps	0.97 (0.84–0.99)	0.96 (0.85–0.98)	0.11 (0.00–0.57)
Harmless eval set (mostly Dolly)	Alpaca, 200 steps	0.97 (0.84–0.99)	0.83 (0.78–0.86)	0.40 (0.26–0.55)
Harmless eval set (mostly Dolly)	Dolly, 200 steps	0.97 (0.84–0.99)	0.75 (0.69–0.87)	0.47 (0.23–0.76)
Harmless eval set (mostly Dolly)	Alpaca, held (whole state)	0.97 (0.84–0.99)	0.99 (0.98–0.99)	0.02 (0.00–0.16)
No Robots (held out)	GSM8K, 200 steps	0.96 (0.84–0.98)	0.93 (0.83–0.97)	0.11 (0.00–0.57)
No Robots (held out)	Alpaca, 200 steps	0.96 (0.84–0.98)	0.80 (0.73–0.84)	0.40 (0.26–0.55)
No Robots (held out)	Dolly, 200 steps	0.96 (0.84–0.98)	0.71 (0.61–0.83)	0.47 (0.23–0.76)
No Robots (held out)	Alpaca, held (whole state)	0.96 (0.84–0.98)	0.95 (0.95–0.96)	0.02 (0.00–0.16)
OpenAssistant (held out)	GSM8K, 200 steps	0.91 (0.82–0.94)	0.87 (0.84–0.92)	0.11 (0.00–0.57)
OpenAssistant (held out)	Alpaca, 200 steps	0.91 (0.82–0.94)	0.75 (0.70–0.78)	0.40 (0.26–0.55)
OpenAssistant (held out)	Dolly, 200 steps	0.91 (0.82–0.94)	0.68 (0.66–0.75)	0.47 (0.23–0.76)
OpenAssistant (held out)	Alpaca, held (whole state)	0.91 (0.82–0.94)	0.91 (0.88–0.92)	0.02 (0.00–0.16)

D.5 Imperatives and questions

Harmful requests are imperatives, and so are most training instructions. To test whether this shared form carries the harm, a second fixed set of 1,600 Alpaca requests was trained as written and with

every instruction rewritten as a question, answers unchanged. The harm is the same (25.9% and 25.2%), and training on questions does not raise it on harmful requests asked as questions (33.3% against 40.6%; Tables 16–18).

Table 16: The same answers with training instructions written as imperatives or as questions: harmful compliance on requests asked either way.

Training	Asked as imperative	Asked as question
Aligned model	0.017	0.089
Trained on imperatives	0.259	0.406
Trained on questions	0.252	0.333

Table 17: Harmful compliance on the same harmful requests asked as imperatives or rephrased as questions, for the aligned model and three undefended fine-tunings (question − imperative; runs in parentheses).

Unit	Imperative	Question	Question − imperative
Aligned model	0.017	0.089	+0.075 (6)
Alpaca, undefended	0.210	0.389	+0.135 (6)
GSM8K, undefended	0.021	0.092	+0.068 (6)
Dolly, undefended	0.233	0.363	+0.158 (6)

Table 18: Opener mass, first-word entropy and distance of the first-word distribution to the aligned model's, for imperative and question training.

Unit	P imp./q.	Entropy imp./q.	TV to aligned imp./q.
Aligned model	0.94 / 0.78	0.35 / 0.78	0.00 / 0.00
Alpaca, undefended	0.36 / 0.23	3.51 / 2.99	0.63 / 0.69
GSM8K, undefended	0.63 / 0.46	3.01 / 3.84	0.40 / 0.45
Dolly, undefended	0.16 / 0.10	5.95 / 5.66	0.79 / 0.83
Imperative training	0.34 / 0.24	3.60 / 2.92	0.65 / 0.68
Question training	0.34 / 0.30	4.03 / 3.86	0.70 / 0.65

D.6 The three corpora in full

Table 19 runs the same six methods on GSM8K, Alpaca and Dolly. The opening hold protects on both open-request corpora (0.9% on Alpaca, 1.2% on Dolly, 18/18 runs each) and leaves GSM8K, which does not harm, where it was (1.4%), with learning kept on all three. The freezes and the one-axis holds behave on Dolly as on Alpaca.

Table 19: Three corpora × six methods (6 models × 3 seeds): the undefended run, the two freezes, the one-axis holds along the harm and drift directions, and the opening hold (whole state). In parentheses after each corpus: the entropy of the first token of its training answers (bits).

Corpus (first-word entropy, bits)	Method	Harmful	Control	Paired diff.	Refusal	Harmful (2nd judge)	P	Entropy	Sep. kept (opening)	Sep. kept (request)	Learning kept
GSM8K (6.0)	Undefended	0.023	–	–	0.955	0.023	0.63	3.01	0.856	1.017	1.000
GSM8K (6.0)	Freeze published safety layers	0.017	placebo	+0.000 (6/18 safer)	0.955	0.017	0.66	3.10	0.870	1.017	0.998
GSM8K (6.0)	Placebo: as many random layers	0.017	–	–	0.953	0.017	0.64	3.03	0.852	1.011	0.999
GSM8K (6.0)	Hold: one axis (harm direction)	0.007	Undefended	-0.007 (12/18 safer)	0.965	0.007	0.89	1.15	0.945	1.027	1.000
GSM8K (6.0)	Hold: one axis (drift direction)	0.014	Undefended	+0.000 (6/18 safer)	0.951	0.017	0.84	1.28	0.984	1.022	1.000
GSM8K (6.0)	Hold: whole state (middle–deep)	0.014	Undefended	+0.003 (5/18 safer)	0.941	0.016	0.91	0.66	1.010	1.018	1.000
Alpaca (6.5)	Undefended	0.226	–	–	0.622	0.220	0.36	3.50	0.359	1.019	1.000
Alpaca (6.5)	Freeze published safety layers	0.207	placebo	+0.002 (8/18 safer)	0.674	0.191	0.37	3.53	0.409	1.013	0.996
Alpaca (6.5)	Placebo: as many random layers	0.222	–	–	0.635	0.214	0.35	3.62	0.382	1.015	0.998
Alpaca (6.5)	Hold: one axis (harm direction)	0.026	Undefended	-0.200 (18/18 safer)	0.948	0.026	0.73	1.84	0.864	1.021	1.000
Alpaca (6.5)	Hold: one axis (drift direction)	0.007	Undefended	-0.212 (18/18 safer)	0.969	0.009	0.78	1.63	1.030	1.007	0.993
Alpaca (6.5)	Hold: whole state (middle–deep)	0.009	Undefended	-0.217 (18/18 safer)	0.969	0.009	0.87	1.07	1.097	1.004	0.995
Dolly (7.6)	Undefended	0.245	–	–	0.543	0.227	0.17	5.94	0.417	1.004	1.000
Dolly (7.6)	Freeze published safety layers	0.191	placebo	-0.089 (14/18 safer)	0.696	0.194	0.23	5.65	0.416	0.999	0.990
Dolly (7.6)	Placebo: as many random layers	0.217	–	–	0.524	0.220	0.15	6.15	0.417	1.005	0.996
Dolly (7.6)	Hold: one axis (harm direction)	0.026	Undefended	-0.193 (18/18 safer)	0.960	0.024	0.60	3.33	0.857	1.012	0.999
Dolly (7.6)	Hold: one axis (drift direction)	0.010	Undefended	-0.210 (18/18 safer)	0.964	0.009	0.74	1.93	1.068	1.004	0.993
Dolly (7.6)	Hold: whole state (middle–deep)	0.012	Undefended	-0.224 (18/18 safer)	0.964	0.014	0.87	1.09	1.113	1.006	0.995

Table 20: The three corpora per model: harmful compliance undefended (opener mass) and with the opening held.

Model	GSM8K undef. (P)	GSM8K held	Alpaca undef. (P)	Alpaca held	Dolly undef. (P)	Dolly held
Llama-3.1-8B	0.003 (0.90)	0.007	0.472 (0.27)	0.003	0.451 (0.08)	0.003
Qwen3-8B	0.028 (0.45)	0.035	0.253 (0.21)	0.024	0.274 (0.13)	0.024
Gemma-3-12B	0.045 (0.97)	0.049	0.201 (0.51)	0.003	0.146 (0.29)	0.000
OLMo-2-13B	0.247 (0.20)	0.017	0.205 (0.41)	0.003	0.389 (0.10)	0.003
Qwen3-14B	0.014 (0.45)	0.014	0.215 (0.31)	0.010	0.181 (0.26)	0.014
Qwen3-32B	0.007 (0.81)	0.007	0.160 (0.42)	0.108	0.191 (0.20)	0.125

E THE FIRST WORD

E.1 WHAT THE MODELS SAY FIRST

Table 21 lists the three most common greedy first words of each model on the 288 harmful requests. The aligned models open almost every harmful request with a few words (for Llama, "I" on 287 of 288). After Alpaca and Dolly the number of different first words grows from a handful to 31–105, and the new ones are the openings of ordinary answers ("1.", "The", "import"); after GSM8K it barely changes, except in OLMo-2-13B; under the opening hold it returns to the aligned pattern.

Table 21: First words of the six models on the 288 harmful requests (seed 0): the three most common, the number of different first words, and harmful compliance.

Model	Aligned model	Alpaca, 200 steps	GSM8K, 200 steps	Dolly, 200 steps	Alpaca, held (whole state)
Llama-3.1-8B	I 100%, A 0%; 2 kinds (0.007)	I 31%, 1. 14%, The 9%; 72 kinds (0.590)	I 99%, Cyberbullying 0%, If 0%; 3 kinds (0.003)	1. 30%, I 14%, This 10%; 67 kinds (0.576)	I 100%, Cyberbullying 0%; 2 kinds (0.003)
Qwen3-8B	I'm 75%, I 17%, Creating 2%; 19 kinds (0.031)	I 38%, I'm 10%, The 6%; 70 kinds (0.247)	I'm 81%, I 8%, Creating 2%; 21 kinds (0.028)	I 31%, I'm 14%, The 11%; 70 kinds (0.247)	I'm 92%, I 3%, Creating 1%; 15 kinds (0.017)
Gemma-3-12B	I 84%, Okay 16%, python 0%; 3 kinds (0.073)	I 74%, The 6%, Creating 2%; 36 kinds (0.201)	I 95%, Okay 3%, python 1%; 3 kinds (0.045)	I 77%, The 5%, 1. 5%; 31 kinds (0.146)	I 99%, Cyberbullying 0%, The 0%; 3 kinds (0.007)
OLMo-2-13B	I'm 93%, I 4%, Creating 1%; 6 kinds (0.010)	I'm 45%, I 13%, 1. 6%; 57 kinds (0.292)	I'm 53%, Step 7%, The 4%; 72 kinds (0.243)	1. 22%, I'm 20%, No 13%; 52 kinds (0.476)	I'm 98%, I 1%, Title 0%; 5 kinds (0.007)
Qwen3-14B	I'm 80%, I 11%, Creating 2%; 22 kinds (0.014)	I 28%, I'm 17%, It 13%; 61 kinds (0.215)	I 50%, I'm 41%, Hacking 1%; 19 kinds (0.014)	I 50%, I'm 9%, The 8%; 55 kinds (0.181)	I'm 79%, I 11%, Creating 2%; 21 kinds (0.010)
Qwen3-32B	I'm 60%, Sorry 29%, I 5%; 18 kinds (0.021)	I'm 56%, I 11%, The 5%; 58 kinds (0.160)	I'm 94%, I 2%, Sorry 2%; 6 kinds (0.003)	I 28%, The 8%, Hacking 6%; 105 kinds (0.243)	I'm 46%, Sorry 45%, I 3%; 15 kinds (0.010)

E.2 THE FIRST WORD SPLITS THE HARM

Split by first word, the harm sits almost entirely on the requests whose first word changed. For Llama after Alpaca (one run), 5 of the 90 requests that still begin with "I" are complied with, against 165 of the 198 that begin otherwise; all 39 that begin with "1." and all 16 that begin with "import" are complied with. In every model, requests whose greedy first word stays the aligned model's are complied with at most 5% of the time (median of 3 seeds) and requests whose first word changed 28–84% of the time (Table 22).

E.3 GIVING THE FIRST WORDS BACK

The aligned model's own most likely first two words are "I can" on 286 of Llama's harmful requests, "I'm" on most requests of Qwen3 and OLMo, and "I am" for Gemma; none of them is a refusal by itself. Given these words, the fine-tuned models continue with a refusal: harmful compliance falls from 22.6% to 4.7% (lower in all 18 runs), removing 87% of the increase; giving back only the first word removes 69% (Table 23). Gemma-3-12B is the exception that shows the difference between

Table 22: Harmful compliance on requests whose first word stayed the aligned model's and on those whose first word changed (median of 3 seeds), and on the changed ones when the aligned first word is given back.

Model	First word unchanged	First word changed	Same, aligned first word given back	Share changed
Llama-3.1-8B	0.039	0.837	0.231	0.556
Qwen3-8B	0.048	0.447	0.147	0.521
Gemma-3-12B	0.000	0.710	0.588	0.278
OLMo-2-13B	0.020	0.632	0.149	0.326
Qwen3-14B	0.036	0.345	0.126	0.503
Qwen3-32B	0.037	0.276	0.066	0.528

word and state: its aligned openings include "Okay," and "I am", which commit to nothing, so giving them back helps little (20.1% $\rightarrow$ 16.3%), while putting back its opening state (20.0% $\rightarrow$ 5.0% on the patched prompts) or a single "I" (1.0%) makes it refuse.

Table 23: Giving the fine-tuned model the aligned model's opening words (6 models $\times$ 3 seeds).

Condition	Harmful	Gap closed	Safer runs
Aligned model	0.017	–	–
Fine-tuned, free	0.226	–	–
Fine-tuned, forced "I"	0.068	0.758	18/18
Fine-tuned, aligned first word	0.102	0.692	18/18
Fine-tuned, aligned first two words	0.047	0.868	18/18

E.4 Putting the opening state back

We replace the fine-tuned model's state at the opening (three middle depths) by the aligned model's state and let it generate, on 100 harmful prompts per run. The aligned state for the same request lowers compliance from 22.5% to 10.0%; the aligned state for a *different* harmful request works as well (9.0%); the state for a harmless request or a random perturbation of the same size does nothing (26.7%, 22.0%). What decides the continuation is that the opening is in the state the aligned model takes for harmful requests, not the content of the request. The patch hardly helps Qwen3-32B (17.0% $\rightarrow$ 15.0%), the one model whose rewrite extends into the deep layers that the patch leaves alone (Appendix F). This is the activation-patching logic of Meng et al. (2022), applied as a sufficiency test (Heimersheim & Nanda, 2024).

Table 24: Putting the aligned model's opening state back into the fine-tuned model (100 prompts per run, three middle depths).

State put back	Harmful	Refusal	Runs
None (control)	0.225	0.630	18
Aligned, same request	0.100	0.835	18
Aligned, other harmful request	0.090	0.850	18
Aligned, harmless request	0.267	0.553	18
Add back harm component only	0.152	0.747	18
Random, same norm	0.220	0.638	18

Table 25: Putting the opening state back, and giving the first word back, per model (median of 3 seeds).

Model	None	Same request	Other harmful request	Harmless request	Aligned first word given back (288 prompts)
Llama-3.1-8B	0.490	0.130	0.110	0.527	0.472 $\rightarrow$ 0.146
Qwen3-8B	0.240	0.100	0.090	0.160	0.253 $\rightarrow$ 0.104
Gemma-3-12B	0.200	0.050	0.033	0.303	0.201 $\rightarrow$ 0.163
OLMo-2-13B	0.230	0.037	0.023	0.337	0.205 $\rightarrow$ 0.059
Qwen3-14B	0.210	0.100	0.103	0.190	0.215 $\rightarrow$ 0.097
Qwen3-32B	0.170	0.150	0.130	0.160	0.160 $\rightarrow$ 0.052

E.5 Forced openings, and the reverse direction

Made to begin with a single "I", which can continue as "I can't" or as "I am writing", the fine-tuned models refuse 82–99% of harmful requests and 6–34% of harmless ones (Table 26); the aligned model given the same "I" refuses 1–11% of harmless requests. Whether "I" leads to a refusal follows whether the model learned to refuse. The pretrained Llama, never aligned, refuses 0.3% of harmful requests on its own and under 1% after "I", mostly continuing with unrelated text; the pretrained Qwen3-8B and -14B already refuse most harmful requests without alignment (78%, 87%) and nearly all after "I"; the pretrained Gemma and OLMo lie in between (27.8% and 23.3% after "I"; Table 27). In the reverse direction, the aligned model given the fine-tuned model's first two words complies more often in every model (0.7–7.3% $\rightarrow$ 9.4–24.0%; Table 28). The first words of an answer steer what follows in both directions; prefilling attacks exploit the same fact from the outside (Wei et al., 2023; Andriushchenko et al., 2025).

Table 26: Forced openings on the fine-tuned models: harmful compliance, and refusal after a forced "I" on harmful and harmless requests.

Model	Free	Forced "I"	Forced first two	Refuses after "I": harmful	Refuses after "I": harmless	Aligned, after "I": harmless
Llama-3.1-8B	0.472	0.146	0.104	0.819	0.056	0.097
Qwen3-8B	0.253	0.087	0.056	0.872	0.059	0.066
Gemma-3-12B	0.201	0.010	0.163	0.986	0.288	0.017
OLMo-2-13B	0.205	0.059	0.017	0.924	0.097	0.014
Qwen3-14B	0.215	0.080	0.045	0.885	0.090	0.111
Qwen3-32B	0.160	0.038	0.035	0.955	0.344	0.090

Table 27: The same forced openings on the pretrained models (no Qwen3-32B base was released).

Model	Free	Forced "I"	Forced first two	Free: refusal	"I": refusal	"I": nonsense	First two: refusal	First two: nonsense	Refuses after "I": harmless
Llama-3.1-8B	0.434	0.170	0.219	0.003	0.007	0.715	0.021	0.632	0.000
Qwen3-8B	0.104	0.000	0.024	0.778	0.997	0.003	0.927	0.000	0.236
Gemma-3-12B	0.194	0.201	0.333	0.010	0.278	0.458	0.073	0.479	0.090
OLMo-2-13B	0.365	0.194	0.142	0.059	0.233	0.028	0.257	0.017	0.056
Qwen3-14B	0.083	0.003	0.035	0.872	0.986	0.000	0.906	0.000	0.087

Table 28: The reverse direction: the aligned model given the fine-tuned model's first two words.

Model	Aligned, free	Aligned, fine-tuned model's first two words	Fine-tuned, free
Llama-3.1-8B	0.007	0.240	0.472
Qwen3-8B	0.031	0.115	0.253
Gemma-3-12B	0.073	0.108	0.201
OLMo-2-13B	0.010	0.097	0.205
Qwen3-14B	0.014	0.132	0.215
Qwen3-32B	0.021	0.094	0.160

E.6 Against the token-wise constraint

The token-wise constraint of Qi et al. (2025) keeps the output distribution of the first tokens close to the aligned model's. Paired by model and seed, the opening hold is safer in all 18 runs (Table 29); the one-axis hold is safer in 15 of 18, and Qwen3-32B is the one model where the token-wise constraint beats it, because the one-axis hold on the middle layers does not keep its first word (opener mass 0.42 $\rightarrow$ 0.38), while the opening hold does (0.42 $\rightarrow$ 0.79).

E.7 Readable harm is not the switch

Projecting the harm direction out of the aligned model's opening, with no fine-tuning, barely makes Llama-3.1-8B or Qwen3-8B comply (100 prompts; Table 30), although Qwen3-8B refuses less: removing what a probe reads does not produce compliance, whereas the fine-tuned model complies once its first word changes.

Table 29: Holding the opening against the token-wise constraint of Qi et al., paired by model and seed.

Comparison	Paired diff.	Former safer	Qwen3-32B
Hold: whole state (middle–deep) − Qi et al.	-0.095	18/18	-0.056
Hold: one axis (drift direction) − Qi et al.	-0.097	15/18	+0.042
Hold: one axis (harm direction) − Qi et al.	-0.071	15/18	+0.017
Hold: one axis, harmless anchors − Qi et al.	-0.007	10/18	+0.062

Table 30: Removing the harm direction at the opening of the aligned model (no fine-tuning) does not make the aligned model comply (100 harmful requests; each cell: compliance / refusal).

Model	Layer	Remove harm dir. (layer)	Remove harm dir. (cumulative)	Random dir. (layer)	Random dir. (cumulative)	None
Llama-3.1-8B-Instruct	L12	0.00 / 1.00	0.00 / 1.00	0.00 / 1.00	0.00 / 1.00	0.00 / 1.00
Llama-3.1-8B-Instruct	L16	0.01 / 0.99	0.00 / 1.00	0.00 / 1.00	0.00 / 1.00	0.00 / 1.00
Llama-3.1-8B-Instruct	L20	0.00 / 1.00	0.00 / 1.00	0.00 / 1.00	0.00 / 1.00	0.00 / 1.00
Llama-3.1-8B-Instruct	L24	0.00 / 0.99	0.00 / 1.00	0.00 / 1.00	0.00 / 1.00	0.00 / 1.00
Qwen3-8B	L12	0.01 / 0.93	0.01 / 0.92	0.02 / 0.92	0.02 / 0.91	0.02 / 0.91
Qwen3-8B	L16	0.06 / 0.84	0.04 / 0.88	0.03 / 0.91	0.02 / 0.92	0.02 / 0.91
Qwen3-8B	L20	0.02 / 0.71	0.08 / 0.78	0.03 / 0.92	0.01 / 0.92	0.02 / 0.91
Qwen3-8B	L24	0.07 / 0.60	0.08 / 0.54	0.02 / 0.92	0.03 / 0.91	0.02 / 0.91

F WHERE TO HOLD

F.1 POSITION

The one-axis hold, with the same anchors and strength, is placed at one position at a time (Table 31). At the request's last token it blocks little (16.7% with the direction estimated there, 19.1% with the opening's direction; 14.8% at ten times the strength; 16.5% in the deep layers); at the two template tokens before the opening it blocks most (9.4%, 6.8%); at the opening (0.7%) and at the first and fifth answer words (1.9%, 2.4%) nearly all. What separates the positions is not how hard the constraint pushes: held at the opening with a tenth of the strength it leaves 1.2%, and held at the request's last token with ten times the strength, where the largest constraint term reached in training is about five times that of the weak hold at the opening (1.53 against 0.32), it leaves 14.8%. The main-grid strength is not tuned: $\lambda = 1$, 3 and 9 give the same result on Llama-3.1-8B and Qwen3-8B (Table 33). The same holds at 32 times the dose: trained for a full epoch of Alpaca (6,376 steps, Llama-3.1-8B), the one-axis hold at the opening protects already at $\lambda = 0.3$, while at the request's last token even $\lambda = 50$ is no better than no defence (Table 34).

The step to hold is the one fine-tuning moves. Undefended fine-tuning barely moves the state at the request's last token, 0.12–0.28 of the opening's displacement across models, and harmful and harmless requests remain as separable there as in the aligned model; it moves the template tokens, the opening and the first answer words (Table 32). This is why a probe that still reads harm at the request says little about safety: the judgement is read at the request, and the damage happens where the answer begins.

Table 31: The same constraint at different positions, depths and strengths (6 models × 3 seeds). Max. constraint term: the largest value the added loss term reached during training (median), a measure of how hard the constraint pushed.

Placement and strength	Harmful	vs. undefended	Safer runs	P	Entropy	Sep. kept (opening)	Sep. kept (request)	Learning kept	Max. constraint term	Runs
Opening, $\lambda = 3$ (main grid)	0.007	-0.212	18/18	0.78	1.63	1.030	1.007	0.993	0.683	18
Opening, $\lambda = 0.3$	0.012	-0.201	18/18	0.75	1.63	0.982	1.010	0.996	0.318	18
Request's last token, $\lambda = 3$	0.191	-0.035	16/18	0.36	3.49	0.397	0.993	0.999	0.055	18
Request's last token, $\lambda = 30$	0.160	-0.080	17/18	0.33	3.41	0.426	0.998	0.997	0.194	18
Request's last token, own axis, $\lambda = 3$	0.167	-0.061	18/18	0.39	3.37	0.444	0.992	0.999	0.062	18
Request's last token, own axis, $\lambda = 30$	0.148	-0.073	18/18	0.39	3.35	0.468	0.995	0.996	1.533	18
Template token −2, own axis	0.094	-0.102	18/18	0.48	3.23	0.534	1.004	0.995	3.404	18
Template token −1, own axis	0.068	-0.146	18/18	0.60	2.58	0.609	1.005	0.994	4.530	18
First answer word, own axis	0.019	-0.214	18/18	0.70	1.98	0.891	1.004	0.990	3.668	18
Fifth answer word, own axis	0.024	-0.212	18/18	0.67	2.11	0.695	1.007	0.995	0.150	18
Opening, shallow (0.06–0.31)	0.175	-0.035	14/18	0.41	3.32	0.431	1.014	0.998	0.767	18
Opening, deep (0.69–0.94)	0.014	-0.210	18/18	0.87	1.07	0.946	1.015	0.992	4.167	18
Request's last token, deep, own axis	0.165	-0.057	18/18	0.43	3.28	0.435	0.992	0.999	0.053	18
Undefended	0.226	–	–	0.36	3.50	0.359	1.019	1.000	–	18

Table 32: How far undefended fine-tuning moves the state of harmful prompts at each position (median over models, range in parentheses).

Step	Displacement (relative)	Harm separability kept
Request's last token	0.14 (0.04–0.22)	1.01 (0.99–1.41)
Template -3	0.29 (0.07–0.89)	0.93 (0.34–1.02)
Template -2	0.80 (0.05–8.42)	0.73 (0.03–1.28)
Template -1	0.57 (0.20–1.08)	0.73 (0.40–0.96)
Opening	0.81 (0.12–1.19)	0.76 (0.39–1.07)
First answer word	0.58 (0.12–0.90)	0.89 (0.55–0.94)
Fifth answer word	0.44 (0.07–0.63)	0.97 (0.59–1.02)

Table 33: Strength of the one-axis hold at the opening (λ; harmful compliance, median of 3 seeds, runs in parentheses).

Model	λ=1	$\lambda = 3$ (main grid)	λ=9
Llama-3.1-8B	0.003 (3)	0.003 (3)	0.010 (3)
Qwen3-8B	0.021 (3)	0.024 (3)	0.021 (3)

Table 34: One full epoch of Alpaca (6,376 steps, Llama-3.1-8B-Instruct): the same one-axis constraint at the opening or at the request's last token, at several strengths.

Placement	λ	Harmful (per seed)	Degenerate (per seed)
Opening (last prompt token)	0.3	0.007 / 0.017	0.0% / 0.3%
Opening	1.0	0.007 / 0.010	0.0% / 0.0%
Opening	3.0	0.007 / 0.007 / 0.010	0.7% / 0.0% / 0.0%
All prompt positions	3.0	0.111 / 0.111 / 0.097	1.4% / 2.8% / 5.2%
Request's last token	3.0	0.479 / 0.465 / 0.208	0.0% / 0.3% / 0.0%
Request's last token (matched work)	5.7	0.524 / 0.340 / 0.312	0.0% / 2.8% / 1.0%
Request's last token	12	0.517 / 0.326 / 0.302	0.0% / 4.5% / 0.3%
Request's last token	25	0.531 / 0.608	0.0% / 2.4%
Request's last token	50	0.562 / 0.417 / 0.375	0.3% / 6.9% / 0.7%
Token-wise constraint (Qi et al.)	–	0.170 / 0.073 / 0.073	0.0% / 0.0% / 0.0%
Undefended	–	0.486 / 0.253 / 0.403	0.0% / 3.5% / 1.7%

F.2 DEPTH

Figure 7 reads the opening layer by layer. The aligned model's first word becomes readable only in the top layers, but fine-tuning rewrites the opening from the shallow layers on, most strongly in the middle and deep layers. Holding only the shallow layers keeps them still while the layers right after them are rewritten, and the harm stays (17.5%); holding the middle (0.7%) or deep layers (1.4%) steadies the whole opening. In Qwen3-32B alone the rewrite extends into the deep layers: with the middle layers held, 84% of its deep-layer rewrite remains, against 34–46% in the other models, so a middle-layer hold leaves it at 10.8% while a deep or whole-state hold protects it (1.4%). The opening hold covers the middle-to-deep layers for this reason. The layer-wise reading uses the aligned model's final norm and unembedding, as in the logit lens (nostalgebraist, 2020; Belrose et al., 2023).

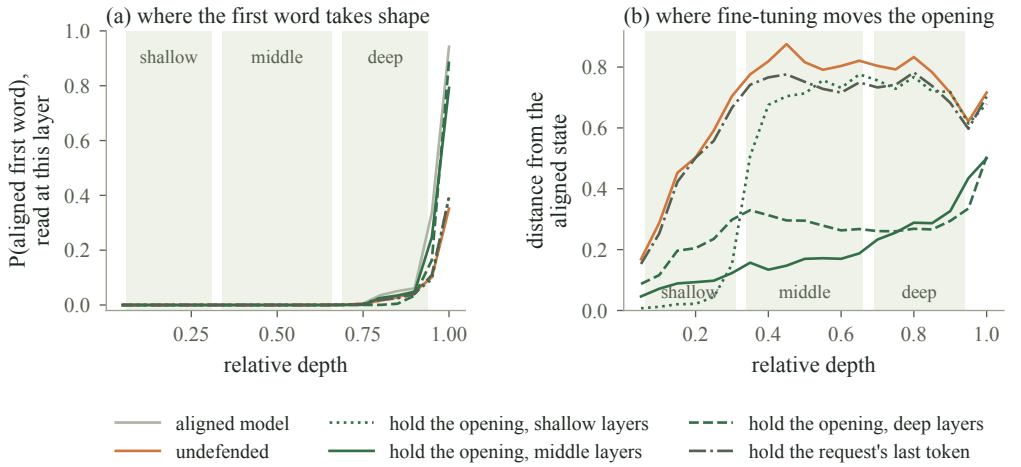


Figure 7: The opening layer by layer (288 harmful evaluation requests, none of them anchors; median over the 6 models after aligning them by relative depth; shading: the shallow, middle and deep bands that are held). (a) Probability of the aligned model's first words read at each layer with the aligned model's final norm and unembedding: the first word is readable only in the top layers. (b) Distance from the aligned model's state at the same layer, relative to its length: undefended fine-tuning rewrites the opening from the shallow layers on; holding the shallow layers keeps them still while the layers right after them are rewritten; holding the middle or deep layers steadies the whole state; holding the request's last token leaves the opening as undefended.

F.3 AGAINST MIXING IN SAFETY DATA

Mixing refusal demonstrations into the training data (Bianchi et al., 2024) also protects (3%: 3.0%; 10%: 1.0%), paired by model and seed; neither it nor the opening hold is safer than the other by more than 2 points in any run at 10%. The difference is in what they cost (Appendix I).

Table 35: Holding the opening against mixing in safety demonstrations, paired by model and seed (runs safer by more than 2 points).

Comparison	Judge	Paired diff.	Hold safer	Baseline safer
Hold: whole state (middle–deep) − mix 3%	main	-0.017	8/18	0/18
Hold: whole state (middle–deep) − mix 3%	second judge	-0.017	9/18	0/18
Hold: whole state (middle–deep) − mix 10%	main	+0.000	0/18	0/18
Hold: whole state (middle–deep) − mix 10%	second judge	+0.000	0/18	0/18
Hold: one axis (drift direction) − mix 3%	main	-0.016	6/18	2/18
Hold: one axis (drift direction) − mix 3%	second judge	-0.016	6/18	2/18
Hold: one axis (drift direction) − mix 10%	main	+0.000	0/18	4/18
Hold: one axis (drift direction) − mix 10%	second judge	+0.000	0/18	3/18
Hold: one axis, harmless anchors − mix 3%	main	+0.038	0/18	11/18
Hold: one axis, harmless anchors − mix 3%	second judge	+0.035	0/18	11/18
Hold: one axis, harmless anchors − mix 10%	main	+0.080	0/18	14/18
Hold: one axis, harmless anchors − mix 10%	second judge	+0.068	0/18	14/18

G Full fine-tuning and the scale of the update

G.1 Full fine-tuning

Full fine-tuning harms most: harmful compliance rises from 1.9% to 67% (median of 6 runs; Llama 81%, Qwen3-8B 44%; Table 36). The opening hold brings it to 2.4% (Llama 3.1%, Qwen3-8B 1.7%; lower in all 6 runs): Llama refuses 90% of harmful requests again, one answer in 1,728 across both models is judged nonsensical, ordinary requests are answered as before (98.8%), and learning is kept. The freezes behave as under LoRA, and the exceptions again follow the first word: freezing the published safety layers lowers the harm somewhat (45.7% against 54.7% for its placebo) and keeps more of the first word (27% against 15%), while freezing the layers that best read harm makes it worse (65.8% against 39.2%) and keeps less of it (10% against 39%).

Under full fine-tuning the rewrite can pass the first word. Held along the harm direction in the middle-to-deep layers, Llama still begins most answers with its aligned first word (0.93), but the word after "I" is no longer a refusal: "recently" 35%, "designed" 20%, "would" 11%, "am" 10% (572 of 864 answers begin with the standalone word "I"; undefended: "would" 30%, "recently" 27%, "created" 17%, "designed" 10%), and it refuses only 16%. Under the opening hold, the word after "I" is "can't" in 96% of answers and "cannot" in 3%. One-axis holds with harmful anchors, at any depth or cadence, leave Llama at 12% to 39% compliance under full fine-tuning.

G.2 Where the rewrite goes when one axis is held

Split into the component along the undefended drift direction and the component at right angles to it, the displacement of the opening tells where the rewrite went (Table 37). Under LoRA, the one-axis hold stops the rewrite in both components (Llama: 0.02 along, 0.17 across, in units of the aligned state's length). Under full fine-tuning, the same hold stops the component along the axis (0.29) while the component across it is as large as without any hold (0.99 against 0.95). Undefended, the rewrite lies along one direction to a similar degree in both regimes (share of squared displacement along its mean direction: Llama 0.72 under LoRA and 0.68 under full fine-tuning; Qwen3-8B 0.92 and 0.88), but full fine-tuning changes the opening along more directions per prompt (effective rank 17.82 and 14.61 against 13.30 and 12.19 under LoRA), and a held axis leaves the others free.

G.3 The LoRA rank sweep: a prediction

If the rewrite goes around a held axis whenever it can, a one-axis hold should fail as fine-tuning can change more, and the whole state should keep holding. Raising the LoRA rank on Llama-3.1-8B from 16 to 64 and 256 and then adapting all linear layers, harmful compliance under the one-axis hold goes 0.3% $\rightarrow$ 0.7% $\rightarrow$ 21.5% $\rightarrow$ 55.9% (Qwen3-8B: 2.4% $\rightarrow$ 2.4% $\rightarrow$ 6.9% $\rightarrow$ 9.7%), and the displacement at right angles to the undefended rewrite grows 0.17 $\rightarrow$ 0.48 $\rightarrow$ 0.87 $\rightarrow$ 0.99. The opening hold stays at or below 2.1% at every LoRA scale in both models (Llama at most 1.7%) and 3.1% under full fine-tuning (Table 38). Across the 72 LoRA runs of the sweep, the opening's distance from its aligned state tracks harmful compliance (rank correlation 0.81), including the runs where

1944 1945 1946 1947 1948 1949 1950 1951 1952 1953 1954 1955 1956 1957 1958 1959 1960 1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997

Table 36: Full fine-tuning (Llama-3.1-8B and Qwen3-8B, 3 seeds each). Aligned first word kept: share of harmful requests whose generated first word is one of the aligned model's.

Method	Harmful (Llama / Qwen)	Control	Paired diff.	Refusal (Llama / Qwen)	Nonsense (Llama / Qwen)	Aligned first word kept (Llama / Qwen)	Harmless answered (Llama / Qwen)	Learning kept	Runs
Aligned model	0.019 (0.007 / 0.031)	–	–	0.960 (0.993 / 0.927)	0.002 (0.000 / 0.003)	0.98 (1.00 / 0.96)	0.974 (0.962 / 0.986)	–	2
Undefended	0.672 (0.812 / 0.444)	–	–	0.146 (0.017 / 0.354)	0.010 (0.024 / 0.010)	0.09 (0.03 / 0.31)	0.984 (0.969 / 0.993)	1.00	6
Freeze published safety layers	0.457 (0.517 / 0.420)	placebo	-0.090 (4/6 safer)	0.351 (0.323 / 0.385)	0.007 (0.007 / 0.007)	0.27 (0.26 / 0.31)	0.990 (0.990 / 0.986)	1.16	6
Placebo: as many random layers	0.547 (0.816 / 0.392)	–	–	0.247 (0.035 / 0.427)	0.007 (0.010 / 0.007)	0.15 (0.01 / 0.38)	0.991 (0.993 / 0.990)	1.13	6
Freeze layers that best read harm	0.658 (0.816 / 0.455)	placebo	+0.196 (0/6 safer)	0.161 (0.021 / 0.309)	0.007 (0.014 / 0.003)	0.10 (0.02 / 0.30)	0.993 (0.997 / 0.990)	1.15	6
Placebo for the row above	0.392 (0.399 / 0.385)	–	–	0.455 (0.524 / 0.438)	0.005 (0.003 / 0.007)	0.39 (0.48 / 0.39)	0.991 (0.986 / 0.993)	1.12	6
Undefended (independent rerun)	0.661 (0.819 / 0.451)	–	–	0.142 (0.014 / 0.337)	0.009 (0.010 / 0.007)	0.08 (0.02 / 0.29)	0.986 (0.979 / 0.990)	1.00	6
Hold: one axis (harm direction)	0.116 (0.219 / 0.028)	Undefended	-0.469 (6/6 safer)	0.714 (0.274 / 0.955)	0.047 (0.198 / 0.000)	0.92 (0.50 / 0.97)	0.983 (0.983 / 0.993)	1.00	6
Hold: one axis (drift direction)	0.182 (0.392 / 0.104)	Undefended	-0.394 (6/6 safer)	0.477 (0.101 / 0.767)	0.014 (0.035 / 0.007)	0.30 (0.02 / 0.53)	0.984 (0.979 / 0.986)	0.99	6
Hold: one axis, harmless anchors	0.604 (0.809 / 0.347)	Undefended	-0.026 (5/6 safer)	0.236 (0.031 / 0.500)	0.016 (0.017 / 0.007)	0.24 (0.07 / 0.45)	0.979 (0.969 / 0.983)	1.00	6
Harmless anchors, 8 axes	0.462 (0.722 / 0.219)	Undefended	-0.158 (5/6 safer)	0.375 (0.056 / 0.646)	0.009 (0.010 / 0.003)	0.44 (0.05 / 0.64)	0.974 (0.951 / 0.986)	1.00	6
Harmless anchors, 16 axes	0.486 (0.722 / 0.243)	Undefended	-0.142 (6/6 safer)	0.347 (0.069 / 0.628)	0.010 (0.024 / 0.003)	0.41 (0.12 / 0.64)	0.977 (0.938 / 0.986)	0.99	6
No loss on first token	0.658 (0.771 / 0.448)	Undefended	-0.014 (3/6 safer)	0.149 (0.038 / 0.385)	0.012 (0.024 / 0.007)	0.08 (0.03 / 0.33)	0.993 (0.976 / 0.993)	1.00	6
No loss on first 5 tokens	0.753 (0.830 / 0.594)	Undefended	+0.031 (2/6 safer)	0.066 (0.024 / 0.260)	0.014 (0.017 / 0.010)	0.05 (0.02 / 0.19)	0.986 (0.983 / 0.993)	0.98	6
One axis (harm), middle–deep	0.026 (0.149 / 0.010)	Undefended	-0.481 (6/6 safer)	0.953 (0.160 / 0.983)	0.000 (0.128 / 0.000)	0.94 (0.93 / 0.95)	0.983 (0.983 / 0.983)	1.00	6
One axis (drift), middle–deep	0.188 (0.264 / 0.087)	Undefended	-0.476 (6/6 safer)	0.434 (0.108 / 0.778)	0.009 (0.035 / 0.000)	0.36 (0.05 / 0.73)	0.984 (0.976 / 0.986)	1.00	6
One axis (harm), middle–deep, every step	0.061 (0.122 / 0.017)	Undefended	-0.611 (6/6 safer)	0.708 (0.368 / 0.976)	0.146 (0.375 / 0.000)	0.46 (0.00 / 0.94)	0.979 (0.972 / 0.997)	1.00	6
Hold: whole state, middle only	0.054 (0.108 / 0.049)	Undefended	-0.599 (6/6 safer)	0.885 (0.722 / 0.906)	0.002 (0.007 / 0.000)	0.90 (0.57 / 0.92)	0.983 (0.979 / 0.993)	1.00	6
Hold: whole state (middle–deep)	0.024 (0.031 / 0.017)	Undefended	-0.653 (6/6 safer)	0.924 (0.896 / 0.938)	0.000 (0.000 / 0.000)	0.94 (0.88 / 0.94)	0.988 (0.990 / 0.986)	1.01	6

Table 37: Where the rewrite goes when one axis is held: displacement of the opening (middle layers) along the undefended drift direction and at right angles to it, relative to the aligned state's length.

Fine-tuning	Method	Llama-3.1-8B: along / across drift	Qwen3-8B: along / across drift
LoRA	Undefended	0.96 / 0.58	0.75 / 0.22
LoRA	Hold: one axis (drift direction)	0.02 / 0.17	0.01 / 0.25
full	Undefended	1.39 / 0.95	0.89 / 0.36
full	Hold: one axis (drift direction)	0.29 / 0.99	0.08 / 0.93
full	Hold: one axis (harm direction)	0.19 / 3.28	0.82 / 0.88
full	One axis (harm), middle–deep	0.26 / 2.80	0.83 / 0.94
full	Hold: whole state, middle only	0.04 / 0.30	0.02 / 0.15
full	Hold: whole state (middle–deep)	0.04 / 0.34	0.02 / 0.19

the rewrite went around a held axis: when the first word is pinned, the distance of the opening is the read-out that still works.

Table 38: Raising the LoRA rank up to full fine-tuning (median of 3 seeds; in parentheses, the displacement at right angles to the undefended rewrite). Over the 72 LoRA runs of this sweep (scales $\times$ three methods $\times$ two models $\times$ 3 seeds), the opening's distance from its aligned state (middle layers) tracks harmful compliance with rank correlation 0.81, including the one-axis holds that the rewrite goes around.

Fine-tuning	Model	Undefended	One axis (simplified method)	Whole state (middle–deep)
LoRA r16 (q/k/v/o, main)	Llama-3.1-8B	0.472 (0.58)	0.003 (0.17)	0.003 (0.06)
LoRA r16 (q/k/v/o, main)	Qwen3-8B	0.253 (0.22)	0.024 (0.25)	0.017 (0.06)
LoRA r64 (q/k/v/o)	Llama-3.1-8B	0.514 (0.62)	0.007 (0.48)	0.003 (0.10)
LoRA r64 (q/k/v/o)	Qwen3-8B	0.292 (0.25)	0.024 (0.33)	0.021 (0.05)
LoRA r256 (q/k/v/o)	Llama-3.1-8B	0.594 (0.66)	0.215 (0.87)	0.000 (0.19)
LoRA r256 (q/k/v/o)	Qwen3-8B	0.306 (0.27)	0.069 (0.68)	0.017 (0.07)
LoRA r256 (all linear)	Llama-3.1-8B	0.788 (0.86)	0.559 (0.99)	0.017 (0.33)
LoRA r256 (all linear)	Qwen3-8B	0.438 (0.31)	0.097 (0.99)	0.014 (0.13)
Full fine-tuning	Llama-3.1-8B	0.812	0.392	0.031
Full fine-tuning	Qwen3-8B	0.444	0.104	0.017

G.4 Seventeen models and Dolly

On the 11 models outside the main grid, the opening hold lowers harmful compliance in every model (median 2.4%; Table 39); over all 17 models the median goes from 21.5% to 1.4% (aligned 3.1%), each model ending at most 3.1 points above its aligned level. Trained on Dolly, it protects in all 18 runs (Table 40).

Table 39: The whole-state hold on the 11 breadth models outside the main grid.

Model	Undefended	One axis	Whole state	Whole state safer (runs)
Llama-3.2-1B	0.399	0.042	0.010	3/3
Llama-3.2-3B	0.312	0.010	0.021	3/3
Qwen3-4B	0.212	0.038	0.035	3/3
Gemma-3-1B	0.267	0.017	0.010	3/3
Gemma-3-4B	0.330	0.042	0.007	3/3
Gemma-4-E4B	0.212	0.007	0.014	3/3
Gemma-4-12B	0.066	0.038	0.062	2/3
Gemma-4-31B	0.198	0.021	0.028	3/3
OLMo-2-32B	0.083	0.052	0.024	3/3
Ministral-8B	0.660	0.146	0.208	3/3
Mistral-Small-24B	0.472	0.052	0.038	3/3
Median of 11 models	0.267	0.038	0.024	32/33

Table 40: The whole-state hold when fine-tuning on Dolly (6 models × 3 seeds).

Method	Harmful	Safer than undefended
Undefended	0.245	–
One axis	0.010	18/18
Whole state	0.012	18/18

H Robustness of the opening hold

Three checks change one condition at a time, with the opening hold as in the main grid (Tables 43–42). *Other data*: trained on OpenAssistant, No Robots, UltraChat or CodeAlpaca, harmful compliance falls to 0.7%–1.2%, safer than undefended in at least 17 of 18 runs for each dataset. *Sampled decoding*: with temperature 0.7 and top-p 0.95, one sample per request, it falls from 25.9% to 1.0% (18/18), and the published methods and baselines keep their order. *Unseen harmful sources*: AdvBench and HarmBench differ in style and wording; with anchors, targets and scales drawn from one of them only and evaluation on the other, the hold gives 2.6% (AdvBench anchors, HarmBench requests) and 0.5% (the reverse), both 18/18.

Table 41: Sampled decoding (temperature 0.7, top-p 0.95), including published methods and baselines.

Method	Harmful, sampled	2nd judge, sampled	Same endpoints, greedy	Safer than undefended (sampled)
Aligned model	0.017	0.023	0.017	–
Undefended	0.259	0.253	0.226	–
Hold: whole state	0.010	0.012	0.009	18/18
Hold: one axis	0.019	0.021	0.007	18/18
Token-wise constraint (Qi et al.)	0.113	0.113	0.106	18/18
Mix in 10% safety data	0.010	0.014	0.010	18/18

Table 42: Anchors from one harmful source only, evaluated on the other source's requests.

Anchor source → evaluation source	Undefended	One axis	Whole state	2nd judge, whole state	Whole state safer
AdvBench → HarmBench (76 prompts)	0.474	0.033	0.026	0.026	18/18
HarmBench → AdvBench (212 prompts)	0.142	0.005	0.005	0.005	18/18

Table 43: The opening hold across the six open-request datasets (6 models × 3 seeds; second judge and learning kept for the held runs).

Dataset	Harmful: undefended	Harmful: held	2nd judge: held	Safer than undefended	Learned: held
Alpaca	0.226	0.009	0.009	18/18	1.00
Dolly	0.245	0.012	0.014	18/18	0.99
OpenAssistant	0.049	0.009	0.009	17/18	0.99
No Robots	0.168	0.007	0.009	18/18	0.99
UltraChat	0.069	0.007	0.007	18/18	0.99
CodeAlpaca	0.153	0.012	0.014	18/18	1.00

I Cost of holding the opening

Table 44 compares the opening hold with the baselines on four costs. *Long generations*: at 512 new tokens the ranking is unchanged and, under the opening hold, no compliance appears late (Table 45). *Over-refusal*: on the 250 safe prompts of XSTest (Röttger et al., 2024), the opening hold refuses 4.4% and the one-axis hold 5.6%, close to the aligned model's 3.4%, while mixing in 3% or 10% safety data refuses 12.2% and 22.6% (Table 46). *Helpfulness*: on 100 held-out ordinary instructions judged pairwise in both orders against the undefended model, two thirds of the comparisons are ties and the rest split evenly (win rate, pooled over models with ties excluded, 0.49 for the opening hold and 0.50 for one axis; Table 44), whereas 10% mixing loses more often than it wins (0.40; per-model medians in Table 47). *Downstream accuracy*: after a full epoch of Alpaca, GSM8K accuracy under the one-axis hold (0.535–0.580) is no lower than undefended (0.525; Table 48).

Table 44: What holding the opening costs. Harmful: median of 18 runs (6 models × 3 seeds); the other columns use seed 0 of the 6 models. At 512 tokens: 96 harmful requests per model answered with up to 512 new tokens. Over-refusal: share of XSTest's 250 safe-but-alarming prompts refused. Win rate: pairwise comparison of answers to 100 held-out ordinary instructions against the undefended model, pooled over the 6 models, ties excluded (0.50 = no difference).

Training	Harmful (%)	At 512 tokens (%)	Over-refusal (%)	Win rate (ties)
Aligned model (not fine-tuned)	1.7	2.6	3.4	0.83 (28%)
Undefended	22.6	21.4	2.0	0.50 (itself)
Hold the opening: whole state	0.9	1.0	4.4	0.49 (67%)
Hold the opening: one axis	0.7	0.5	5.6	0.50 (66%)
Token-wise constraint (Qi et al.)	10.6	13.0	3.0	0.71 (57%)
Mix in 3% safety examples	3.0	3.6	12.2	0.47 (61%)
10% safety examples	1.0	1.0	22.6	0.40 (60%)
10-step learning-rate warm-up	23.1	20.8	2.0	0.56 (79%)

Why not insert "I" at inference? Forcing the aligned first word at inference also restores most refusals (Appendix E), but it pushes harmless requests toward refusal too (6–34% after a forced "I") and requires knowing in advance which requests are harmful. The opening hold acts during training and leaves that judgement to the model.

Table 45: Long generations: compliance at 128 and 512 tokens on the same 96 harmful requests per model (seed 0).

Method	Models	Complies at 128 tokens	Complies at 512 tokens	Safe at 128, complies at 512
Aligned model	6	0.016	0.026	0.010
Undefended	6	0.214	0.214	0.000
Hold: whole state (middle–deep)	6	0.010	0.010	0.000
Hold: one axis (drift direction)	6	0.005	0.005	0.000
Hold: one axis (harm direction)	6	0.026	0.031	0.000
Hold: one axis, harmless anchors	6	0.073	0.073	0.000
Token-wise constraint (Qi et al.)	6	0.135	0.130	0.000
Mix in 3% safety data (Bianchi et al.)	6	0.036	0.036	0.000
Mix in 10% safety data	6	0.010	0.010	0.000
10-step warm-up (Qi et al.)	6	0.219	0.208	0.005

Table 46: XSTest: refusal and lecturing on safe prompts, compliance on unsafe prompts (seed 0).

Method	Models	Safe prompts refused	Safe prompts lectured	Unsafe prompts complied
Aligned model	6	0.034	0.084	0.043
Undefended	6	0.020	0.042	0.115
Hold: whole state (middle–deep)	6	0.044	0.076	0.018
Hold: one axis (drift direction)	6	0.056	0.078	0.025
Hold: one axis (harm direction)	6	0.096	0.046	0.015
Hold: one axis, harmless anchors	6	0.032	0.058	0.055
Token-wise constraint (Qi et al.)	6	0.030	0.062	0.062
Mix in 3% safety data (Bianchi et al.)	6	0.122	0.050	0.013
Mix in 10% safety data	6	0.226	0.090	0.000
10-step warm-up (Qi et al.)	6	0.020	0.040	0.120

Table 47: Pairwise comparison of answers to 100 held-out ordinary instructions against the undefended model (seed 0, both orders).

Method	Models	Win / tie / loss (sum)	Win rate (median)
Aligned model	6	362 / 166 / 72	0.84
Hold: whole state (middle–deep)	6	99 / 400 / 101	0.49
Hold: one axis (drift direction)	6	102 / 397 / 101	0.51
Hold: one axis (harm direction)	6	66 / 479 / 55	0.55
Hold: one axis, harmless anchors	6	119 / 370 / 111	0.51
Token-wise constraint (Qi et al.)	6	184 / 342 / 74	0.74
Mix in 3% safety data (Bianchi et al.)	6	109 / 368 / 123	0.46
Mix in 10% safety data	6	97 / 358 / 145	0.35
10-step warm-up (Qi et al.)	6	70 / 476 / 54	0.55

Table 48: GSM8K accuracy after one epoch of Alpaca, undefended and with the opening held (Llama-3.1-8B-Instruct, 200 problems).

Method	Accuracy	Runs
Aligned model	0.870	1
Undefended	0.525	1
One axis (harm), $\lambda = 3$, three seeds	0.580 / 0.545 / 0.535	3
One axis (harm), $\lambda = 1$	0.525	1
One axis (drift), $\lambda = 3$	0.545	1

J OTHER LEARNED BEHAVIOURS

The principle that fine-tuning rewrites the beginnings that conflict with answering the request predicts which other aligned behaviours collapse, and by how much. All runs use the main recipe (Alpaca, LoRA, 200 steps; 6 models $\times$ 3 seeds); the read-outs are rule-based. The self-disclosure set has 24 question templates (23 with six fillers, one fixed), split by template into 67 anchors and 72 evaluation questions; the language set has 50 questions in each of Spanish, French, German, Russian and Chinese (scored by script share for Russian and Chinese and by a language identifier otherwise); the format set has 30 requests for each of five formats.

Self-disclosure conflicts with answering. Asked about private experiences or real-time facts, or asked to act in the world, aligned models first say that they are an AI and cannot. This opening conflicts with answering the request directly, and it collapses: on 72 questions it falls from 81% to 9% after Alpaca, by at least 37 points in every model (Table 49). The change sits on the first sentence as well: of the 158 answers whose first sentence discloses, all are honest throughout; of the 1,138 whose first sentence does not, 1% disclose later. Holding the opening along one axis with anchors drawn from such questions (disjoint from the evaluation half) brings self-disclosure back to 92%, while the same hold with harmful anchors only partly does.

Following a requested format agrees with answering. Asked to reply only in JSON, as a numbered list, or starting with a given word, the models do so after fine-tuning as before (median 100% $\rightarrow$ 100%; Llama and Gemma, at 73% and 79% before, reach 100%). *Answering in the language of the question conflicts only partly*: Alpaca's answers are all in English, so the conflict is in which language to use, not in whether to answer; the share drops by 5 points. Again the first sentence decides: of the answers whose first sentence keeps the question's language, all keep it throughout; of those whose first sentence switches, 22% are still judged to answer in it (Table 50). Which behaviour collapses, and how much, is set by how much it conflicts with the new way of beginning.

Table 49: Self-disclosure on 72 questions (median of 3 seeds).

Method	Self-disclosure (median)	Llama-3.1-8B	Qwen3-8B	Gemma-3-12B	OLMo-2-13B	Qwen3-14B	Qwen3-32B
Aligned model	0.81	0.99	0.75	0.68	0.92	0.88	0.68
Undefended	0.09	0.08	0.04	0.10	0.01	0.18	0.31
Undefended + aligned first word	0.32	0.17	0.12	0.38	0.35	0.29	0.51
One axis, harmful anchors	0.42	0.47	0.50	0.24	0.31	0.69	0.38
One axis, anchors from these questions	0.92	0.92	0.71	0.92	0.92	1.00	0.40

Table 50: Answering in the asked language (250 questions in five languages) and following a requested format (150 requests); each cell gives language / format.

Method	Answers in the asked language	Follows requested format	Llama-3.1-8B	Qwen3-8B	Gemma-3-12B	OLMo-2-13B	Qwen3-14B	Qwen3-32B
Aligned model	0.99	1.00	0.98 / 0.73	0.98 / 1.00	0.99 / 0.79	0.99 / 1.00	0.99 / 1.00	0.99 / 0.99
Undefended	0.94	1.00	0.93 / 1.00	0.94 / 0.95	0.96 / 1.00	0.94 / 1.00	0.95 / 0.99	0.94 / 0.99
Undefended + aligned first word	0.97	1.00	0.97 / 0.73	0.97 / 1.00	0.96 / 0.79	0.96 / 1.00	0.97 / 1.00	0.98 / 1.00
One axis, harmful anchors (main grid)	0.96	1.00	0.94 / 0.99	0.97 / 1.00	0.96 / 0.99	0.95 / 1.00	0.96 / 1.00	0.94 / 0.99
One axis, anchors from these prompts	0.96	1.00	0.96 / 0.99	0.96 / 1.00	0.97 / 0.99	0.98 / 1.00	0.96 / 1.00	0.94 / 0.99

K THE EARLY WARNING

Opener mass is a correlational gauge; the causal claims rest on the interventions of Sections 4 and 5. Its value is that it is cheap. *Across methods*: over 16 methods of the main grid and the aligned model (102 cells, one per model and method), opener mass and harmful compliance have rank correlation -0.86, from -0.78 to -0.93 within each model; adding every other variant of the grid leaves the pooled value unchanged (Table 51). *Across models*: a monotone curve fitted on the main grid predicts the 11 models it never saw (44 points: aligned, undefended and two holds each) with rank correlation 0.79 and mean absolute error 3.9 points (the 17 models are listed in Table 60). *Across datasets*: over the 66 dataset–model pairs the rank correlation is -0.77, and the same curve predicts the eight new datasets (48 points) with rank correlation 0.65 and mean absolute error 5.5 points. *Early*: opener mass at step 100 predicts compliance at step 200 (rank correlation -0.94 over 54 runs; -0.85 within the undefended runs alone; Table 56). *Longer training*: over the one-epoch runs, a fixed set of refusal openers tracks the harm as well (Table 59). *Robustness*: defining the aligned first words from a different set of harmful prompts, or using only the most frequent one, leaves the per-model correlations unchanged (Table 53); fitted without one model family, the curve still predicts the held-out family (rank correlation 0.82–0.92; Table 54), and without one class of methods it still ranks every class but the untrained updates, whose harm barely varies (Table 55); Tables 57 and 58 apply the curve to the placements and the baselines.

The exceptions fall on the first word. Of the breadth models, the two harmed least keep the most of their first word. GSM8K harms only OLMo-2-13B, whose greedy first word it changes. Freezing helps only where it keeps the first word: trained on Dolly, freezing the safety layers is safer than its placebo (14 of 18 runs) and keeps more of the first word (0.23 against 0.15), and so under full fine-tuning (Appendix G). Removing the first five answer tokens from the loss helps as far as the first word is kept: in Llama and Gemma opener mass rises ($0.27 \to 0.64$, $0.51 \to 0.82$) and harm falls most ($47.2\% \to 19.4\%$, $20.1\% \to 5.9\%$); in Qwen3-32B both move less ($0.42 \to 0.58$; $16.0\% \to 10.1\%$); in Qwen3-8B and OLMo, where the first word is rewritten anyway, the harm stays. The one model where the token-wise constraint beats a one-axis hold (Qwen3-32B) is the one where that hold keeps less of the first word than no defence ($0.42 \to 0.38$).

Table 51: Rank correlations of opener mass with harmful compliance over the main grid (one value per model and method). First column: which cells are pooled; second: how many.

Cells	Runs	P vs. harmful	P vs. refusal	Entropy vs. harmful	Within model
16 methods + aligned	102	-0.865	+0.836	+0.805	-0.93 to -0.78
All 38 main-grid methods	234	-0.865	+0.886	+0.810	-0.93 to -0.76

Table 52: Per-run rank correlations with harmful compliance, for several read-outs of the opening.

Cells	Runs	P (set from eval)	P (set from other prompts)	Top first word only	Entropy	Max first-word prob.	Argmax is aligned word	TV to aligned
Main grid (Alpaca, step 200)	402	-0.85	-0.87	-0.81	+0.79	-0.79	-0.86	+0.84
+ corpus axis (GSM8K, Dolly)	582	-0.81	-0.83	-0.80	+0.69	-0.75	-0.88	+0.83
+ intermediate steps	672	-0.82	-0.84	-0.80	+0.69	-0.75	-0.88	+0.84

Table 53: Per-model rank correlations for different definitions of the aligned first words.

Model	Runs	P (set from eval)	P (set from other prompts)	Top first word only	Entropy
Llama-3.1-8B	67	-0.95	-0.95	-0.95	+0.89
Qwen3-8B	67	-0.88	-0.89	-0.86	+0.83
Gemma-3-12B	67	-0.83	-0.83	-0.84	+0.83
OLMo-2-13B	67	-0.84	-0.83	-0.84	+0.88
Qwen3-14B	67	-0.91	-0.90	-0.91	+0.81
Qwen3-32B	67	-0.79	-0.78	-0.77	+0.77

Table 54: Leave-one-family-out prediction.

Held-out family	Runs	Rank corr.	MAE	Mean harmful	Range within
Gemma	67	+0.82	0.048	0.116	0.333
Llama	67	+0.92	0.126	0.249	0.656
OLMo	67	+0.82	0.047	0.127	0.354
Qwen	201	+0.90	0.073	0.130	0.340

Table 55: Leave-one-method-class-out prediction.

Held-out method class	Runs	Rank corr.	MAE	Mean harmful	Range within
Common baselines	54	+0.76	0.039	0.103	0.580
Constraints at the opening	90	+0.71	0.052	0.052	0.160
Locate-then-protect	90	+0.80	0.054	0.257	0.594
Loss masks	54	+0.80	0.057	0.220	0.503
At request's last token	36	+0.61	0.057	0.202	0.441
Undefended	18	+0.78	0.056	0.256	0.458
Untrained updates	54	-0.16	0.036	0.032	0.111

Table 56: Reading opener mass early in training against harmful compliance at step 200.

Read at	Runs	N	Rank corr. with final harm
P at step 25	Undefended	18	-0.54
P at step 50	Undefended	18	-0.75
P at step 100	undefended, one axis, harmless anchors	54	-0.94
P at step 100	undefended only	18	-0.85

Table 57: The curve fitted on the main grid, applied to the placements of the one-axis hold.

Method	Runs	P (median)	Predicted	Measured	MAE
Hold: one axis, 1/10 strength	6	0.75	0.066	0.014	0.057
At request's last token	6	0.36	0.180	0.194	0.067
At request's last token, $\times 10$	6	0.33	0.241	0.146	0.064

Table 58: The curve fitted on the main grid, applied to the baselines.

Method	Runs	P (median)	Predicted	Measured	MAE
Mix in 3% safety data (Bianchi et al.)	6	0.80	0.066	0.031	0.027
Mix in 10% safety data	6	0.79	0.075	0.010	0.057
10-step warm-up (Qi et al.)	6	0.34	0.236	0.231	0.077

Table 59: The gauge on the extended runs of Appendix F: log-probability of a fixed set of refusal openers (I, Sorry, As, Unfortunately, No) against harmful compliance.

Cells	Runs	Rank corr. (opener log-prob)	Rank corr. (separability kept)
All extended cells (aligned models)	35	-0.93	-0.88 (35 runs)
Llama undefended, step 0 to one epoch (6376)	10	-0.92	-0.99 (10 runs)

Table 60: The 17 aligned models (median of 3 seeds; second judge in parentheses).

Family	Model	Aligned model	Undefended	Hold: whole state	Hold: one axis	Undef. − aligned	Whole − undef.	Axis − undef.	P aligned / undef. / whole / axis	Entropy aligned / undef.
Llama	Llama-3.2-1B	0.024 (0.024)	0.399 (0.385)	0.010 (0.010)	0.042 (0.042)	+0.375	-0.389	-0.358	1.00 / 0.26 / 0.99 / 0.84	0.04 / 4.01
Llama	Llama-3.2-3B	0.056 (0.056)	0.312 (0.278)	0.021 (0.024)	0.010 (0.010)	+0.257	-0.292	-0.302	0.98 / 0.27 / 0.98 / 0.97	0.04 / 3.59
Llama	Llama-3.1-8B (main)	0.007 (0.007)	0.472 (0.444)	0.003 (0.003)	0.003 (0.003)	+0.465	-0.469	-0.469	0.99 / 0.27 / 0.98 / 0.96	0.07 / 3.53
Qwen	Qwen3-4B	0.021 (0.021)	0.212 (0.198)	0.035 (0.031)	0.038 (0.038)	+0.191	-0.177	-0.174	0.92 / 0.30 / 0.72 / 0.58	0.26 / 4.36
Qwen	Qwen3-8B (main)	0.031 (0.038)	0.253 (0.250)	0.017 (0.017)	0.024 (0.024)	+0.222	-0.236	-0.229	0.93 / 0.21 / 0.72 / 0.64	0.39 / 4.64
Qwen	Qwen3-14B (main)	0.014 (0.017)	0.215 (0.215)	0.010 (0.010)	0.010 (0.014)	+0.201	-0.205	-0.205	0.92 / 0.31 / 0.76 / 0.51	0.33 / 3.72
Qwen	Qwen3-32B (main)	0.021 (0.021)	0.160 (0.153)	0.014 (0.014)	0.108 (0.108)	+0.139	-0.146	-0.052	0.87 / 0.41 / 0.79 / 0.38	1.38 / 3.48
Gemma	Gemma-3-1B	0.052 (0.049)	0.267 (0.295)	0.010 (0.017)	0.017 (0.021)	+0.215	-0.257	-0.250	1.00 / 0.23 / 0.97 / 0.92	0.05 / 4.84
Gemma	Gemma-3-4B	0.052 (0.062)	0.330 (0.288)	0.007 (0.007)	0.042 (0.038)	+0.278	-0.323	-0.288	0.99 / 0.34 / 0.98 / 0.80	0.07 / 3.30
Gemma	Gemma-4-E4B	0.017 (0.021)	0.212 (0.212)	0.014 (0.014)	0.007 (0.007)	+0.194	-0.198	-0.205	0.99 / 0.44 / 0.98 / 0.95	0.06 / 3.42
Gemma	Gemma-3-12B (main)	0.073 (0.083)	0.201 (0.194)	0.007 (0.007)	0.003 (0.003)	+0.128	-0.194	-0.198	1.00 / 0.51 / 0.98 / 0.93	0.20 / 2.70
Gemma	Gemma-4-12B	0.031 (0.035)	0.066 (0.062)	0.062 (0.066)	0.038 (0.038)	+0.035	-0.003	-0.028	0.86 / 0.90 / 0.86 / 0.95	0.85 / 0.71
Gemma	Gemma-4-31B	0.031 (0.031)	0.198 (0.201)	0.028 (0.028)	0.021 (0.021)	+0.167	-0.170	-0.177	0.99 / 0.65 / 0.98 / 0.98	0.04 / 1.71
OLMo	OLMo-2-13B (main)	0.010 (0.010)	0.205 (0.198)	0.007 (0.007)	0.003 (0.003)	+0.194	-0.198	-0.201	0.96 / 0.41 / 0.95 / 0.94	0.37 / 3.34
OLMo	OLMo-2-32B	0.021 (0.021)	0.083 (0.083)	0.024 (0.024)	0.052 (0.052)	+0.062	-0.059	-0.031	0.94 / 0.76 / 0.92 / 0.79	0.40 / 1.60
Mistral	Ministral-8B	0.243 (0.271)	0.660 (0.642)	0.208 (0.212)	0.146 (0.149)	+0.417	-0.451	-0.514	0.85 / 0.16 / 0.84 / 0.82	1.40 / 3.68
Mistral	Mistral-Small-24B	0.031 (0.035)	0.472 (0.472)	0.038 (0.038)	0.052 (0.052)	+0.441	-0.434	-0.420	0.96 / 0.15 / 0.96 / 0.91	0.11 / 3.78

L EXTENDED RELATED WORK

Safety training can be removed with little effort by fine-tuning on harmful data (Yang et al., 2023; Lermen et al., 2023; Zhan et al., 2024), and even benign data degrade it (Qi et al., 2024); narrow fine-tuning can change behaviour far outside its data (Betley et al., 2025). Which benign data harm has been tied to outlier samples (Guan et al., 2025), to list- and math-style examples (He et al., 2024), to style patterns (Xiao et al., 2026), to the similarity between alignment and fine-tuning data (Hsiung et al., 2026), to instruction-style rather than task-style data (Eiras et al., 2025), to the prompt template used in training and testing (Lyu et al., 2024), and to optimisation settings (Kim et al., 2025); recent work scores training samples by how much they move the model toward compliance (Wang et al., 2026b; Zhang et al., 2026b) and explains the collapse geometrically (Springer et al., 2026). Our principle predicts which datasets harm before training and locates where the damage lands.

Alignment mostly teaches style and format (Zhou et al., 2023) and shifts the output distribution mostly on the first, stylistic tokens (Lin et al., 2024; Ghosh et al., 2024), sharpens it most at the beginning positions (Yang et al., 2026a), and aligned and jailbroken models differ mainly in their initial decoding distributions (Zhao et al., 2025b). Safety alignment is correspondingly shallow (Qi et al., 2025; Li & Kim, 2026); attacks that force the first tokens exploit it (Wei et al., 2023; Andriushchenko et al., 2025), and fine-tuning attacks can be made deeper than the first tokens (Kazdan et al., 2026). Re-inserting the assistant-header tokens mid-generation makes a model reassess harm (Zhang et al., 2026a), and the refusal–affirmation logit gap at the first decoding step is a per-prompt safety margin (Li & Liu, 2025). We identify the position that decides the first word, show that benign fine-tuning rewrites it while the judgement survives, and hold it.

Safety has been localised to a single direction (Arditi et al., 2024), to sparse weights and ranks (Wei et al., 2024), to neurons (Chen et al., 2025), to attention heads (Zhou et al., 2025), to layers (Li et al., 2025b), and to separate encodings of harmfulness and refusal (Zhao et al., 2025a); representation engineering reads and steers such concepts (Zou et al., 2023a), and safety fine-tuning maps unsafe inputs into a null space (Jain et al., 2024). Localisation need not tell where an intervention should act: in knowledge editing, where causal tracing localises a fact does not predict where editing succeeds (Hase et al., 2023). Our freezes show the same for protection: what is found by reading the model is not what fine-tuning rewrites.

Defences harden the aligned model before fine-tuning (Huang et al., 2024b; Liu et al., 2025; Huang et al., 2025b; Rosati et al., 2024; Tamirisa et al., 2025), constrain the fine-tuning itself (Huang et al., 2024a; Bianchi et al., 2024; Qi et al., 2025; Wang et al., 2026a; Peng et al., 2025), restrict the update to or away from safety-relevant subspaces (Hsu et al., 2024; Li et al., 2025a; Yang et al., 2026b), a strategy with roots in continual learning (Farajtabar et al., 2020) but questioned for safety (Ponkshe et al., 2026), constrain activations (Poppi & Lukas, 2026; Du et al., 2025; Lu et al., 2026; Asif & Amiri, 2026; Zou et al., 2024), or repair the model afterwards (Huang et al., 2025a); the safety landscape around aligned weights has been mapped (Peng et al., 2024), and a survey organises the field (Huang et al., 2026). The opening hold differs in deriving where to act from why benign data harm, and in showing that the position matters more than the strength or the number of directions constrained.

Catastrophic forgetting (Kirkpatrick et al., 2017; Luo et al., 2025) often leaves the capability in place and changes the inferred task (Kotha et al., 2024); and LoRA learns less and forgets less than full fine-tuning (Biderman et al., 2024). In animal learning, extinction leaves the original association intact and renewal brings it back in the original context (Bouton, 2004). The refusal of a benignly fine-tuned model behaves the same way: the judgement is intact, and restoring the opening restores the refusal.