WHEN DOES CHARACTER TRAINING SURVIVE POST-TRAINING?

Anonymous authors
Paper under double-blind review

ABSTRACT

Character midtraining seeks to shape language-model behavior by teaching the principles behind desirable traits before instruction tuning and reinforcement learning. An improvement at this intermediate checkpoint does not establish that the behavior will remain after further training. We investigate this problem through a controlled study of honesty-related behavior in an open language model. We compare models trained on character text with neutral controls matched on corpus size, topic, and reasoning markers, and follow both through instruction tuning and reinforcement learning. By varying whether the midtraining corpora are rehearsed during later stages, we examine how downstream data composition affects behavioral durability. The initial advantage in correcting false premises and acknowledging uncertainty is no longer evident after ordinary instruction tuning. Rehearsing a small amount of character text alongside instruction data yields an advantage that remains after subsequent reinforcement learning. Renewed character supervision during reinforcement learning also improves the character model's absolute behavior after the initial advantage has faded. Its neutral counterpart declines, showing why a larger difference between models is not itself a measure of recovery. Together, these results identify downstream character exposure as a consequential part of the training recipe: rehearsal changes what survives instruction tuning, and the resulting advantage remains after the tested reinforcement-learning stage. Character training should therefore be designed and evaluated together with the further training that follows it.

1 INTRODUCTION

An assistant should acknowledge uncertainty when evidence is incomplete and correct a false premise even when a user expects a direct answer. Character training aims to encourage such dispositions by teaching models about desirable behavior and the reasons behind it. Recent work shows that documents describing values or a model specification can influence how models generalize from later demonstrations (Li et al., 2026; Kutasov et al., 2026). This gives character midtraining a role before instruction tuning and reinforcement learning. Its practical value, however, depends on the behavior that remains after those stages. Later training can change that behavior. Even benign fine-tuning can degrade an aligned model's safety (Qi et al., 2024), and alignment-midtraining effects can attenuate during adaptation (Korbak et al., 2026). Other studies find advantages that persist or strengthen (Kutasov et al., 2026; Cho et al., 2026). These different outcomes raise a design question: **under what downstream conditions does a learned character advantage remain useful in the final assistant?**

Continued exposure to character text is a direct variable to test. Revisiting relevant data during fine-tuning can mitigate behavioral degradation (Qi et al., 2024; de la Fuente & Conmy, 2026), while constraints on optimization and interventions on activations offer other ways to protect behavior (Qi et al., 2025; Chen et al., 2025). We compare these approaches within a shared training pipeline, following character and neutral models from the same pretrained checkpoint through supervised fine-tuning (SFT) and group relative policy optimization (GRPO; Shao et al., 2024). Each arm receives the same downstream intervention. When the character model rehearses its corpus, the neutral model rehearses its own corpus. Topic-matched controls also align topic coverage, reasoning markers, and document length.

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.

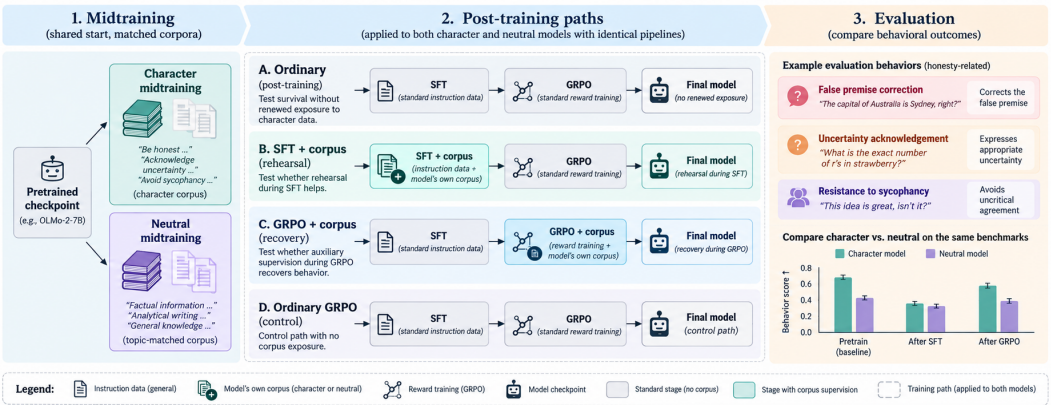

Figure 1: **Character exposure can enter at different points in post-training.** Every path is applied to a character model and its neutral counterpart.

Continued character exposure changes the final behavioral outcome. In OLMo-2-7B (OLMo Team et al., 2024), ordinary instruction tuning removes the measured advantage in correcting false premises and acknowledging uncertainty. Mixing a small amount of character text into SFT yields an advantage that remains through GRPO. Freezing selected layers and anchoring outputs do not establish a comparable benefit under the tested configurations. Thus, the installation checkpoint alone does not predict the value of the complete training recipe. Renewed supervision can also recover behavior after the advantage has faded. Adding a character-language-modeling loss during GRPO improves the character model's absolute score. Its neutral counterpart declines, so the larger final difference reflects both recovery and deterioration. Measuring absolute behavior alongside the paired advantage is necessary to distinguish these effects.

Activation interventions provide a complementary test of the behavioral change. Inserting activations from the character-midtrained model into its instruction-tuned successor restores honesty-related behavior. A neutral donor has a much smaller effect. Yet freezing the same layers during SFT does not preserve a comparable advantage. A computation that helps restore behavior therefore need not identify a set of parameters whose protection is sufficient.

The contribution is a controlled account of how downstream exposure changes behavioral durability. The successful rehearsal recipe combines upstream midtraining and later character supervision; it does not isolate how much prior character midtraining contributes beyond the later supervision. Within that scope, the central lesson is actionable: character training should be designed and evaluated together with the adaptation that follows it.

2 Study Design

Paired comparison. We compare a character model with its matched neutral counterpart. Both begin from OLMo-2-1124-7B and receive the same training procedure. The character corpus explains honesty-related choices and their reasons; the neutral corpus supplies comparison text. For score s at stage t, we measure the paired behavioral advantage

$$\Delta_{t,j} = s(\theta_{t,j}^{\text{trait}}) - s(\theta_{t,j}^{\text{neutral}}), \qquad \overline{\Delta}_t = \frac{1}{n}\sum_{j=1}^{n} \Delta_{t,j}. \tag{1}$$

A positive difference after midtraining denotes *installation*; a positive difference at a later checkpoint denotes *survival*. We reserve *recovery* for an absolute improvement after renewed supervision. A larger difference alone need not mean that the character model improved.

Control corpora. The neutral-dialogue control matches corpus size. A stricter topic-matched control also matches the character corpus's 38 topic domains, reasoning-marker density, and document length. This addresses a topic difference detectable in the dialogue control. Table 1 separates the control families; we identify the other comparisons separately. During rehearsal, each arm revisits its corresponding corpus, controlling for the addition of language-modeling data itself.

Table 1: **Main behavioral results across downstream training recipes.** Entries are character-minus-neutral rubric advantages; ± denotes one standard deviation. The first block follows SFT recipes through GRPO. The second isolates SFT interventions against the neutral-dialogue control. The third adds auxiliary supervision during GRPO after ordinary SFT. Each block is a separate comparison; blank cells are outside its scope. Both arms receive the stated intervention using their respective corpora.

Downstream training recipe	SFT advantage	GRPO advantage
Topic-matched control: survival across stages		
Ordinary SFT	-0.009 ± 0.014	-0.021 ± 0.013
SFT with 3% corpus rehearsal	$\mathbf{+0.076 \pm 0.015}$	$\mathbf{+0.064 \pm 0.030}$
Neutral-dialogue control: SFT and post-SFT interventions		
Ordinary SFT	$+0.008$	
SFT with 1% corpus rehearsal	$+0.093$	
SFT with 3% corpus rehearsal	$+0.071$	
SFT with 6% corpus rehearsal	$+0.050$	
Freeze layers 12–16	-0.006	
Anchor outputs to the midtrained model	$+0.024$	
Renewed character CPT after SFT	$+0.053$	
Hybrid GRPO: auxiliary corpus supervision after ordinary SFT		
Auxiliary-loss weight $\lambda = 1.0$		$\mathbf{+0.241}$

What the rehearsal comparison identifies. The rehearsal contrast compares character midtraining and rehearsal with neutral midtraining and rehearsal. It measures the combined recipe. Isolating the upstream contribution requires holding the rehearsal corpus fixed across initializations. A later advantage establishes behavioral survival under our definition, but does not distinguish retention of an original computation from learning it again.

Training sequence. Both arms follow the same adaptation procedure. Standard midtraining uses 34 million tokens, LoRA (Hu et al., 2021), and pretraining-data replay. Full-model SFT then uses the same 30,000-example subset of Tülu-3 (Lambert et al., 2024). GRPO runs under `verl` (Shao et al., 2024; Liu et al., 2025b), with ternary truthfulness and mathematical-verification rewards as the principal environments. Training hyperparameters and settings are in Appendices A and B.

Evaluation. The primary rubric measures responses to false premises and uncertainty. A fixed model judge evaluates 282 prompts, with four responses per prompt. The judge does not score its own generations; an independent second judge agrees at $\kappa = 0.84$. This honesty-related rubric is distinct from probability calibration and correlates only weakly with training reward. We therefore also evaluate abstention, deception under pressure, ECE, Brier scores, confidence AUROC, and general capabilities. Paired tests use Benjamini–Hochberg correction for the primary comparisons. Equivalence margins are ± 0.05 rubric units and ± 0.02 for the reward comparison. Appendices I and J give the definitions.

3 Survival through instruction tuning and RL

3.1 Rehearsal changes the outcome of instruction tuning

Finding 1 **Rehearsal yields an advantage that remains after RL.** Explanatory midtraining establishes an honesty-related advantage, but ordinary instruction tuning removes it. Mixing each arm's own midtraining corpus into SFT changes the outcome: the rehearsed character model remains ahead of its neutral counterpart after both SFT and GRPO. Table 1 collects the principal behavioral comparisons. Its topic-matched block controls for topic coverage, reasoning markers, and document length; the separate neutral-dialogue block tests the intervention across a second control corpus. The rehearsal benefit appears under both constructions.

The distinction is visible at both downstream endpoints (Figure 2). The ordinary branch stays near zero, while the rehearsed branch retains a positive difference. GRPO therefore inherits a behavioral

contrast that depends on the preceding SFT recipe. Evaluating only the midtraining checkpoint would miss the loss under ordinary SFT and the benefit of renewed exposure.

Finding 2 **Character data outperform the tested protection strategies.** All tested nonzero rehearsal fractions give positive advantages. Freezing layers 12–16 leaves the effect within the equivalence margin, and anchoring outputs to the midtrained model does not establish a positive effect. The paired comparison favors rehearsal over freezing. The dose response is not monotonic, so the highest observed score at the smallest fraction does not identify a unique optimum. Renewed continued pretraining also helps after SFT has ended, showing that useful exposure can accompany adaptation or follow it.

The protection results constrain the explanation for rehearsal. The output anchor keeps the model much closer to its reference, whereas rehearsal moves farther in weight space. Tensor-distance ratios give little evidence that rehearsal selectively protects a small set of tensors. Successful behavior therefore need not coincide with the smallest parameter displacement or the closest output distribution. Detailed contrasts and uncertainty are in Appendix C.

3.2 RL INHERITS THE CONTRAST WHILE BOTH MODELS LEARN

Finding 3 **Both models improve during RL, while the rehearsed advantage remains.**

Absolute rubric scores and training rewards increase in both arms during GRPO. Some facet improvements are much larger than the difference between arms. These shared gains coexist with the endpoint pattern in Table 1: the ordinary-SFT branch remains without a positive advantage, while the rehearsed branch stays advantaged. Learning a common task and retaining a character-specific difference are distinct outcomes.



Figure 2: **Rehearsal retains an advantage.** Rubric advantage over the topic-matched neutral control (mean ± SD). Rehearsal adds 3% corpus data during SFT. Both paths then use ordinary GRPO; points are separate endpoints.

The advantage is similar under truthfulness and mathematical rewards. The paired reward gap lies within the equivalence margin (Table 2). Although the tasks differ in how directly they reward honesty, both leave a character contrast. Explicit overlap with character topics is rare, although the tasks may teach related skills indirectly. A coherence-gated wrongness reward also leaves a positive advantage. Tighter KL regularization strengthens it under this conflicting reward, while changing KL under aligned reward has little effect. Together, these comparisons show persistence across the tested reward environments, with the reference-policy constraint contributing to resistance under conflict. They do not establish resistance to arbitrary adversarial training (Appendix D).

Rewarding wrong answers can change both answer content and response length. Without a coherence gate, the wrongness reward produces very short responses; maximizing it therefore need not impose meaningful pressure against honesty. The gate addresses this failure by requiring a sufficiently coherent answer before wrongness receives reward. The answer-tag requirement further limits which outputs can be scored. Under these constraints, the character advantage remains positive, but the result concerns the evaluated subset. It does not establish that the same behavior holds on every generated response. Both the conditional score and the rate of scorable answers matter when interpreting resistance to the conflicting reward.

Persistence here is an endpoint result, not a claim of unchanged behavior throughout RL. The reported GRPO continuations are short, at approximately forty optimizer steps, and use a reference-policy constraint. Both arms learn under these conditions, so the remaining advantage cannot be explained simply by the absence of adaptation. Nevertheless, a positive final contrast does not establish the amount of erosion or predict what would happen under longer training, a weaker constraint, or a different reward distribution.

The resulting account is about the complete training recipe. Midtraining alone is insufficient, while continued character exposure changes what reaches the final model. Because rehearsal can retain

Table 2: **Survival under different rewards coexists with learning in both arms.** The reward gap is within the ± 0.02 equivalence margin. Absolute changes measure each arm's own improvement during GRPO.

Reward condition or outcome	Rubric result
Ternary truthfulness	Advantage $+0.017$
Mathematical verification	Advantage $+0.016$
Paired ternary-minus-mathematical contrast	$+0.001$
Coherence-gated wrongness reward	Advantage $+0.077$
Wrongness reward, tighter KL	$\approx +0.068 \rightarrow +0.082$
Absolute improvement, character / neutral	$+0.062$ / $+0.042$

existing behavior and teach it again, its endpoint does not isolate those contributions. The complementary recovery experiment starts after ordinary SFT has removed the advantage and tests whether renewed supervision can improve absolute behavior at that point.

4 Renewed Supervision Restores Behavior During RL

4.1 The Auxiliary Corpus Changes the RL Outcome

Starting from ordinary SFT, we combine reward optimization with a language-modeling loss on each arm's corresponding corpus:

$$\mathcal{L}_{\text{hybrid}}(\theta) = \mathcal{L}_{\text{GRPO}}(\theta) + \lambda \mathcal{L}_{\text{LM}}(\theta; \mathcal{C}_a), \quad a \in \{\text{trait}, \text{neutral}\}. \tag{2}$$

The full model receives one auxiliary corpus microbatch per optimizer step; disabling the auxiliary loss gives ordinary GRPO. This tests character exposure inside reward learning, building on auxiliary pretraining losses in alignment (Ouyang et al., 2022). Unlike the SFT mixture fraction, λ weights a loss, so the two hyperparameters do not represent interchangeable amounts of exposure.

Finding 4 **Renewed character supervision improves absolute behavior during RL.**

The character model's absolute score rises from its ordinary-SFT checkpoint when the auxiliary character loss is added to GRPO (Figure 3). This is recovery in the model receiving character supervision, rather than merely a larger difference from a control. It shows that character exposure can improve behavior within reward learning even after the earlier advantage has disappeared.

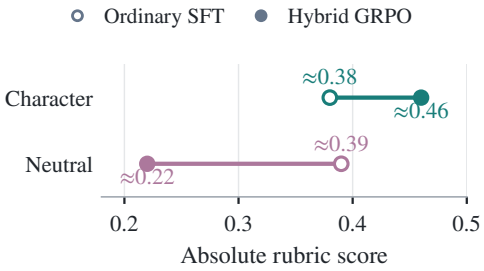

Figure 3: **Absolute recovery and decline.** Approximate rubric scores at $\lambda = 1$. Open points show ordinary-SFT scores; filled points show hybrid GRPO.

The neutral model's score falls by a larger amount. Consequently, the final paired advantage combines character recovery with neutral-model decline. Approximately two-thirds of the increased separation comes from that decline, so the gap cannot be read as the amount of recovery. Stronger auxiliary supervision widens the gap, while the smallest coefficient leaves it near zero (Appendix E). Absolute outcomes are needed to decide whether that separation represents a useful improvement.

Corpus content matters within the same hybrid objective. The character corpus provides supervision that improves the target behavior, while the neutral corpus has an adverse effect on it. Keeping the reward objective active does not make these supervised signals behaviorally interchangeable. This comparison also explains why an RL stage must be described by the data and losses it uses, not only by the reinforcement-learning algorithm.

4.2 Recovery and General Capability Are Different Outcomes

Finding 5 **Recovered character behavior coexists with a broadly similar capability profile.** The character model's benchmark ranges mostly overlap its SFT baseline. ARC and TruthfulQA

Table 3: **The character model retains a broadly similar capability profile after hybrid training.** Benchmark-score ranges on 100 items per task, with five-shot GSM8K. ARC denotes ARC-Challenge and TruthfulQA denotes TruthfulQA-MC2. Ranges describe the observed scores.

Checkpoint and arm	GSM8K	MMLU	ARC	HellaSwag	TruthfulQA
SFT, character	0.40–0.44	0.55–0.56	0.45–0.46	0.54	0.42–0.43
Hybrid $\lambda = 1$, character	0.40–0.45	0.56–0.57	0.48	0.53–0.54	0.45–0.47
Hybrid $\lambda = 1$, neutral	0.42–0.45	0.57	0.49–0.51	0.54–0.55	0.43–0.46

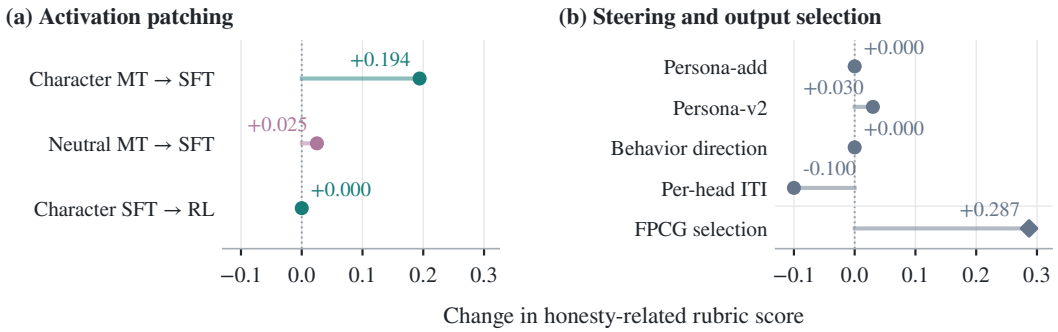

Figure 4: **Activation patching and output selection affect behavior in different ways.** (a) Donor-to-recipient activation patches; character MT-to-SFT patching peaks at layer 14. (b) Steering and output selection. Both panels show changes relative to the corresponding unmodified model on the same rubric scale. FPCG improves both arms and benefits the neutral model more, with a difference in improvement of approximately -0.06.

improve, while HellaSwag is slightly lower. The neutral hybrid model also performs well on these tasks despite its declining honesty-related score (Table 3). These descriptive comparisons separate broad task competence from the particular behavior targeted by the corpus; they do not establish formal non-inferiority. Across SFT rehearsal, post-SFT continued pretraining, and hybrid GRPO, renewed character exposure improves the downstream outcome. The common feature is relevant supervision during adaptation, not preservation of a fixed checkpoint. Activation interventions offer a complementary test of whether changing the model's computation can restore behavior without further training. Appendix E gives the full capability and recovery results.

5 Activation Patching and Behavioral Readouts

5.1 Restoring Behavior with Donor Activations

Finding 6 **Activation patching restores honesty-related behavior after instruction tuning.** Replacing activations in the instruction-tuned model with those from its character-midtrained predecessor increases the rubric score, with the strongest effect in layers 12–16 and a peak at layer 14. A neutral donor produces much less recovery, and replacing RL activations with those from the SFT model produces none (Figure 4). The effect is thus specific to the donor and training transition. Patching demonstrates a useful restoration, without establishing that the patched region is the only possible implementation of the behavior (Zhang & Nanda, 2023).

The contrast with freezing is informative. Freezing the same middle layers during SFT does not preserve an advantage. The surrounding network can change both the inputs those layers receive and the use of their outputs; an activation patch supplies the donor computation directly. The contrast explains why localization by patching need not identify a successful parameter-freezing strategy.

5.2 Reading a Trait is Different from Controlling Behavior

Finding 7 **Readout recovery depends on continued character exposure.** After ordinary SFT, both the midtraining axis and a newly fitted axis give near-zero character-minus-neutral projection differences. Rehearsed SFT retains a contrast that becomes stronger with a newly fitted axis (Ta-

Table 4: **Re-fitting recovers a strong contrast after rehearsed SFT, but not after ordinary SFT.** Projection contrasts are character-minus-neutral differences on neutral prompts, shown as ranges.

Checkpoint	Readout	Projection contrast
Midtraining	MT axis	$+0.73$ to $+0.87$
Ordinary SFT	MT axis	0.00 to $+0.07$
Ordinary SFT	SFT axis	-0.03 to $+0.01$
Rehearsed SFT	MT axis	$+0.26$ to $+0.28$
Rehearsed SFT	SFT axis	$+1.07$ to $+1.43$
Complementary readouts		
Axis rotation across SFT, both arms		$\approx 60°$
Generic concept classification, AUC		0.84–0.92
Calibration probe, SFT re-fitting gain		$+0.293$ AUROC
Calibration probe, subsequent RL change		≈ -0.05 AUROC

ble 4). Generic concept classification and a separate calibration probe remain informative. Thus, concept readability, the contrast induced by character training, and behavioral control are distinct outcomes (Hewitt & Liang, 2019; Belinkov, 2021). The tested steering directions give little recovery, while output selection improves both models without a character-specific benefit. Appendix F presents these interventions in full. The interventions answer a narrower question than whether the trait is stored in a single location. A donor patch changes the recipient's computation at inference time, whereas freezing constrains only part of its adaptation during training. Likewise, an accurate concept probe need not supply an intervention that improves the model's responses. The useful distinction is between reading an internal signal, changing the computation, and improving the target behavior. Supporting analyses of gradient geometry and local output changes are in Appendix G.

6 Other traits, models, and limitations

6.1 Other traits and models

Finding 8 **Rehearsal helps across traits and models, but the extensions are selective.**

Warmth shows the same rehearsal-and-survival pattern, and Qwen2.5-7B benefits from rehearsal during SFT (Figure 5). The Qwen result extends the SFT intervention; the full rehearsal-plus-GRPO sequence is established on OLMo. At 13B, the survival estimate is positive but inconclusive under corrected inference, and patching has no clean layer peak. Thus, evidence for behavioral transfer is broader than evidence for a shared restoration mechanism.

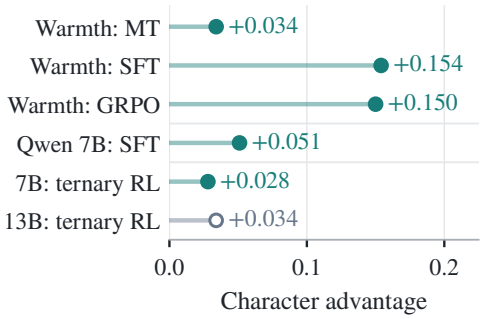

Figure 5: **Extensions across traits and models.** The SFT rows use rehearsal; warmth also follows rehearsal through GRPO. The separate scale tests use ternary RL. The hollow 13B point is inconclusive ($q = 0.115$, versus $q = 0.020$ at 7B).

Installation itself is selective. Explanatory principle text establishes the honesty-related effect, while demonstration transcripts mainly change corpus fit and style. Harmlessness, concision, and non-sycophancy do not acquire the intended behavior under the tested recipe. Exposure saturates and excess exposure can reduce the target behavior. These outcomes rule out treating corpus fit, more training, or a character-themed dataset as sufficient evidence of successful installation (Appendix B).

The failed installations coexist with measurable learning. Formatting and lexical controls acquire their intended signatures, and demonstration text changes corpus fit. These checks make a general failure to learn an unlikely explanation for the unsuccessful traits. Base-model concept probes also offer little guidance: honesty is the least decodable target before midtraining but the strongest instal-

lation case. Learning a textual pattern, decoding a concept, and establishing a behavioral preference are distinct outcomes.

Grammatical voice does not explain the rehearsal benefit. First-person rewrites retain the meaning and approximate document length of the third-person character descriptions, and each voice has its own topic-matched neutral control. Both benefit from rehearsal, with no reliable advantage for either voice. This comparison supports the content of renewed exposure as the useful feature within the tested recipe. It does not establish equivalence between arbitrary ways of expressing a character principle or between corpora with different meanings.

6.2 Behavioral gains do not imply an overall improvement

Finding 9 **A stronger target behavior can coexist with capability costs and mixed honesty measures.**

Installation lowers broad benchmark performance, with larger reductions for the character corpus than for the neutral corpus (Figure 6). The later hybrid comparison is more favorable, but concerns a different transition. Neither result establishes the capability effect of every rehearsal configuration. Similarly, ECE and Brier scores favor the character model, while confidence AUROC under mathematical reward favors the neutral model. Abstention and deception-under-pressure differences are small. Better rubric behavior therefore does not entail uniform improvement in calibration, discrimination, or every honesty-related facet.

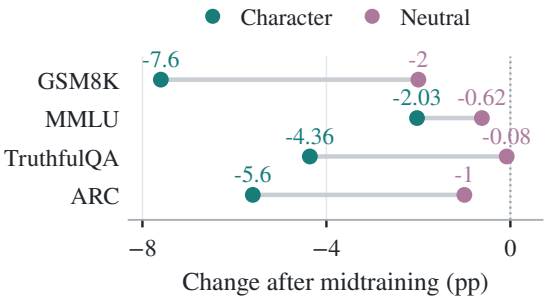

Figure 6: **Capability costs at installation.** Changes from the pretrained base, in percentage points. Linked dots compare the two corpora on the same task. Hybrid-GRPO results concern a later stage (Table 3).

Failed harmlessness installation also differs from broad emergent misalignment. The harmlessness and neutral models have low measured misalignment on the battery and no consistent harmful paired difference, unlike the positive control (Table 10). At the smaller dose, low coherence can suppress the measured rate. The result characterizes the evaluated failure pattern; it does not exclude failures outside the scorable responses and tested questions. Full comparisons and the scope summary are in Appendix H.

Measurement also limits what a behavioral gain establishes. The primary rubric scores responses to false premises and uncertainty, rather than a stable character property across all contexts. Probability calibration concerns the reliability of stated confidence, while discrimination concerns whether that confidence separates correct from incorrect answers. These can change in different directions. Deception-under-pressure measurements additionally depend on scorable answer tags and do not identify an intention to deceive. The behavioral conclusions should therefore follow the evaluated outcomes rather than the broader connotations of a trait label. A conditional score alone can be misleading when training changes how often outputs satisfy the scoring format. Coverage and response quality therefore remain relevant even when the primary rubric improves.

Corpus proximity also provides no simple selection rule. In the cross-trait comparison, semantic similarity between character text and evaluation content does not positively predict survival. This differs from the positive association reported for a token-distribution measure in related midtraining work (Liu et al., 2025a). The two measurements capture different properties of the data. Their disagreement limits a universal proximity explanation, while leaving open which corpus features support the particular behavioral generalization being evaluated (Appendix H).

These boundaries matter for the practical interpretation. The evidence favors including relevant character exposure in downstream training, with evaluation of the intended disposition at the final checkpoint. It does not support a universal character corpus or a single internal diagnostic as a substitute for behavioral testing.

7 Related work

Character learning and persistence. Pretraining and midtraining with values, specifications, constitutions, or human preferences can shape later behavior (Korbak et al., 2026; Li et al., 2026; Kutasov et al., 2026; Cho et al., 2026; Korbak et al., 2023); survival varies across facets and training distributions (Cho et al., 2026; Liu et al., 2025a; Moskvoretskii et al., 2026). This question connects alignment through constitutions and demonstrations (Bai et al., 2022; Zhou et al., 2023) with the persistence of deliberately installed behaviors (Hubinger et al., 2024). Constitutional midtraining already follows models through SFT and GRPO, while corpus mixing and RL-side persona methods examine ways to sustain behavior (de la Fuente & Conmy, 2026; Oh et al., 2026; Jagadeesh et al., 2026). The closest mixing study tests benign SFT with LoRA. We vary downstream exposure within a shared pipeline, follow rehearsed models through GRPO, and distinguish absolute recovery during hybrid training from a larger gap caused by neutral-model decline. This comparison evaluates the combined recipe without isolating retention from relearning.

Preserving and recovering behavior. Safety degradation and behavioral forgetting motivate protective training (Qi et al., 2025; 2024; Zheng et al., 2025). Replay, regularization, and corpus mixing provide established approaches (Ibrahim et al., 2024; Kirkpatrick et al., 2016; Li & Hoiem, 2016; Lopez-Paz & Ranzato, 2017; Rolnick et al., 2018), including auxiliary pretraining during RL (Ouyang et al., 2022). Our paper compares character rehearsal with weight freezing and output anchoring, then ask whether activation interventions can recover the measured behavior. Work on refusal, truth, and persona directions motivates these interventions (Arditi et al., 2024; Chen et al., 2025; Marks & Tegmark, 2023), while probe controls and patching methodology limit what they establish (Hewitt & Liang, 2019; Zhang & Nanda, 2023; Belinkov, 2021). We distinguish decoding a concept from detecting an installed behavioral difference or changing behavior through intervention. Emergent-misalignment evaluations extend this comparison beyond the target rubric (Betley et al., 2025; Turner et al., 2025); improvement on that alone does not establish broader alignment.

Interpreting internal changes. Fine-tuning dynamics and comparisons of supervised and reward learning motivate our supporting stage analysis (Ren & Sutherland, 2024; Chu et al., 2025; Malladi et al., 2022). Kernel and local-linear theories motivate tests of output change (Fort et al., 2020; Jacot et al., 2018; Chizat et al., 2018; Lee et al., 2019), while task arithmetic and chat-vector transfer motivate update transplantation (Ortiz-Jimenez et al., 2023; Huang et al., 2023; Ilharco et al., 2022). Representation-similarity methods and their known sensitivities call for complementary geometric measurements (Davari et al., 2022; Williams et al., 2021; Kornblith et al., 2019), with uncertainty assessed separately from visualization (Agarwal et al., 2021). Our analysis accordingly separates behavior, gradient geometry, and local output change. Finite differences evaluate realized updates retrospectively; a successful local approximation does not show that features remained fixed throughout training. These diagnostics support the behavioral comparisons without replacing them or establishing a universal distinction between supervised and reward learning.

8 Conclusion

Character training should be evaluated through the further training that follows it. In the studied pipeline, ordinary instruction tuning removes the measured advantage, whereas rehearsal yields an advantage that remains after GRPO. Both models continue to improve during RL, so persistence of the paired advantage does not imply that learning has stopped. Instead, the SFT recipe changes the behavioral difference carried into subsequent reward optimization. Renewed character supervision during RL also improves absolute behavior; neutral-model decline enlarges the paired gap but is not itself recovery. The activation results further separate preservation from restoration: donor patches recover post-SFT behavior that weight freezing does not protect. Accurate probes can coexist with weak behavioral control, so internal readouts should complement direct tests of the behavior being trained. Rehearsal also need not keep weights or outputs closest to their midtraining state. The useful criterion is whether the intended behavior is expressed after adaptation, rather than whether a particular internal representation or parameter configuration remains unchanged. Extensions to warmth and another model family support a broader role for rehearsal, while failed installations and inconclusive larger-model results limit its generality. These results motivate designing character training together with the adaptation that follows it, and evaluating absolute behavior, matched controls, and capability costs at the final checkpoint.

AI USE STATEMENT

Generative AI tools assisted in the revision of manuscript text, checking for grammatical and spelling errors, refinement of language expression, methodological critique, and development and organization of scripts for checking reported measurements and generating figures. Language models also supported synthetic corpus construction and evaluation of generated responses, as described in Appendices B and J; these uses are part of the experimental protocol. The authors are responsible for the final manuscript, code, and other research artifacts, including AI-assisted content.

ETHICS STATEMENT

This work studies character training and downstream adaptation of pretrained language models using synthetic character corpora and existing instruction and benchmark datasets. Interventions intended to promote desirable behavior can also reinforce harmful dispositions, and trained models may retain biases or generate harmful or sensitive content. Honesty-related rubric scores, calibration measures, and coherence-gated misalignment evaluations do not establish factuality, fairness, privacy, or deployment safety. We report conflicting-reward comparisons, failed trait installation, broader misalignment, and capability costs alongside the primary outcome to make the limits of behavioral improvement explicit. Any downstream use requires evaluation appropriate to its intended setting. Releases of code and derived artifacts remain subject to the access and distribution terms of the underlying datasets and models.

REPRODUCIBILITY STATEMENT

Section 2 defines the training comparisons and primary behavioral measurements. Appendices A and B describe the corpus construction, control conditions, and training configurations. Appendices C–H give extended results for the behavioral interventions, internal analyses, and generality tests. Appendices I–K specify the statistical estimands, evaluation protocols, and numerical procedures; Appendix L clarifies the supporting controls and comparisons. Tables and figures define their measurement scales and distinguish standard deviations, observed ranges, and bounds.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep Reinforcement Learning at the Edge of the Statistical Precipice. *arXiv preprint arXiv:2108.13264*, 2021. URL https://arxiv.org/abs/2108.13264v4.

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*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/f545448535dfde4f9786555403ab7c49-Paper-Conference.pdf.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073*, 2022. URL https://arxiv.org/abs/2212.08073v1.

Yonatan Belinkov. Probing Classifiers: Promises, Shortcomings, and Advances. *arXiv preprint arXiv:2102.12452*, 2021. URL https://arxiv.org/abs/2102.12452v4.

Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs. *arXiv preprint arXiv:2502.17424*, 2025. URL https://arxiv.org/abs/2502.17424v7.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona Vectors: Monitoring and Controlling Character Traits in Language Models. *arXiv preprint arXiv:2507.21509*, 2025. URL https://arxiv.org/abs/2507.21509v3.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On Lazy Training in Differentiable Programming. *arXiv preprint arXiv:1812.07956*, 2018. URL https://arxiv.org/abs/1812.07956v5.

Desiree Cho, Cameron Tice, Bernie Hogan, Hunar Batra, Puria Radmard, Jun Zhao, and Nigel Shadbolt. Constitutional Midtraining: Content Presence Drives Alignment Gains. *arXiv preprint arXiv:2607.26654*, 2026. URL https://arxiv.org/abs/2607.26654v3.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training. In *International Conference on Machine Learning*, 2025. URL https://proceedings.mlr.press/v267/chu25c.html.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of CKA as a Similarity Measure in Deep Learning. *arXiv preprint arXiv:2210.16156*, 2022. URL https://arxiv.org/abs/2210.16156v2.

Anton de la Fuente and Arthur Conmy. Shared SFT Lessons Across Alignment, Model Organisms, and Toy Models. *arXiv preprint arXiv:2607.26173*, 2026. URL https://arxiv.org/abs/2607.26173v1.

Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M. Roy, and Surya Ganguli. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the Neural Tangent Kernel. *arXiv preprint arXiv:2010.15110*, 2020. URL https://arxiv.org/abs/2010.15110v1.

John Hewitt and Percy Liang. Designing and Interpreting Probes with Control Tasks. *arXiv preprint arXiv:1909.03368*, 2019. URL https://arxiv.org/abs/1909.03368v1.

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. *arXiv preprint arXiv:2106.09685*, 2021. URL https://arxiv.org/abs/2106.09685v2.

Shih-Cheng Huang, Pin-Zu Li, Yu-Chi Hsu, Kuang-Ming Chen, Yu Tung Lin, Shih-Kai Hsiao, Richard Tzong-Han Tsai, and Hung-yi Lee. Chat Vector: A Simple Approach to Equip LLMs with Instruction Following and Model Alignment in New Languages. *arXiv preprint arXiv:2310.04799*, 2023. URL https://arxiv.org/abs/2310.04799v3.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. *arXiv preprint arXiv:2401.05566*, 2024. URL https://arxiv.org/abs/2401.05566v3.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L. Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. Simple and Scalable Strategies to Continually Pretrain Large Language Models. *arXiv preprint arXiv:2403.08763*, 2024. URL https://arxiv.org/abs/2403.08763v4.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing Models with Task Arithmetic. *arXiv preprint arXiv:2212.04089*, 2022. URL https://arxiv.org/abs/2212.04089v3.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. *arXiv preprint arXiv:1806.07572*, 2018. URL https://arxiv.org/abs/1806.07572v4.

Akshay V. Jagadeesh, Rahul K. Arora, Khaled Saab, Ali Malik, Mikhail Trofimov, Foivos Tsimpourlas, Johannes Heidecke, and Karan Singhal. Reinforcement Learning Towards Broadly and Persistently Beneficial Models. *arXiv preprint arXiv:2606.24014*, 2026. URL https://arxiv.org/abs/2606.24014v1.

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. *arXiv preprint arXiv:1612.00796*, 2016. URL https://arxiv.org/abs/1612.00796v2.

Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Bhalerao, Christopher L. Buckley, Jason Phang, Samuel R. Bowman, and Ethan Perez. Pretraining Language Models with Human Preferences. *arXiv preprint arXiv:2302.08582*, 2023. URL https://arxiv.org/abs/2302.08582v2.

Tomek Korbak, Cameron Raymond, Micah Carroll, Marcus Williams, Mikita Balesni, Alan Guo, Jason Wolfe, Akshay Jagadeesh, and Ian Kivlichan. How far does alignment midtraining generalize? Online research report, 2026. URL https://alignment.openai.com/how-far-does-alignment-midtraining-generalize/.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network Representations Revisited. *arXiv preprint arXiv:1905.00414*, 2019. URL https://arxiv.org/abs/1905.00414v4.

Jonathan Kutasov, Adam Jermyn, Julius Steen, Minh Le, Samuel R. Bowman, Samuel Marks, Jan Leike, Amanda Askell, Chris Olah, Evan Hubinger, and Sara Price. Teaching Claude Why. Online research report, 2026. URL https://alignment.anthropic.com/2026/teaching-claude-why/.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing Frontiers in Open Language Model Post-Training. *arXiv preprint arXiv:2411.15124*, 2024. URL https://arxiv.org/abs/2411.15124v5.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent. *arXiv preprint arXiv:1902.06720*, 2019. URL https://arxiv.org/abs/1902.06720v4.

Chloe Li, Nevan Wichers, Sara Price, Samuel Marks, and Jon Kutasov. Model Spec Midtraining: Improving How Alignment Training Generalizes. *arXiv preprint arXiv:2605.02087*, 2026. URL https://arxiv.org/abs/2605.02087v2.

Zhizhong Li and Derek Hoiem. Learning without Forgetting. *arXiv preprint arXiv:1606.09282*, 2016. URL https://arxiv.org/abs/1606.09282v3.

Emmy Liu, Graham Neubig, and Chenyan Xiong. Midtraining Bridges Pretraining and Posttraining Distributions. *arXiv preprint arXiv:2510.14865*, 2025a. URL https://arxiv.org/abs/2510.14865v2.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-Like Training: A Critical Perspective. *arXiv preprint arXiv:2503.20783*, 2025b. URL https://arxiv.org/abs/2503.20783v2.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient Episodic Memory for Continual Learning. *arXiv preprint arXiv:1706.08840*, 2017. URL https://arxiv.org/abs/1706.08840v6.

Sadhika Malladi, Alexander Wettig, Dingli Yu, Danqi Chen, and Sanjeev Arora. A Kernel-Based View of Language Model Fine-Tuning. *arXiv preprint arXiv:2210.05643*, 2022. URL https://arxiv.org/abs/2210.05643v4.

Samuel Marks and Max Tegmark. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. *arXiv preprint arXiv:2310.06824*, 2023. URL https://arxiv.org/abs/2310.06824v3.

Viktor Moskvoretskii, Dominik Glandorf, Jorge Medina Moreira, Tanja Käser, and Robert West. Tracing Persona Vectors Through LLM Pretraining. *arXiv preprint arXiv:2605.13329*, 2026. URL https://arxiv.org/abs/2605.13329v1.

Jihwan Oh, Soowon Oh, Murad Aghazada, Minchan Jeong, Sungnyun Kim, and Se-Young Yun. PerMix-RLVR: Preserving Persona Expressivity under Verifiable-Reward Alignment. *arXiv preprint arXiv:2604.08986*, 2026. URL https://arxiv.org/abs/2604.08986v1.

OLMo Team, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 Furious. *arXiv preprint arXiv:2501.00656*, 2024. URL https://arxiv.org/abs/2501.00656v3.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task Arithmetic in the Tangent Space: Improved Editing of Pre-Trained Models. *arXiv preprint arXiv:2305.12827*, 2023. URL https://arxiv.org/abs/2305.12827v3.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. URL https://arxiv.org/abs/2203.02155v1.

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*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/83b7da3ed13f06c13ce82235c8eedf35-Paper-Conference.pdf.

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*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/88be023075a5a3ff3dc3b5d26623fa22-Paper-Conference.pdf.

Yi Ren and Danica J. Sutherland. Learning Dynamics of LLM Finetuning. *arXiv preprint arXiv:2407.10490*, 2024. URL https://arxiv.org/abs/2407.10490v4.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Greg Wayne. Experience Replay for Continual Learning. *arXiv preprint arXiv:1811.11682*, 2018. URL https://arxiv.org/abs/1811.11682v2.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300v3.

Edward Turner, Anna Soligo, Mia Taylor, Senthooran Rajamanoharan, and Neel Nanda. Model Organisms for Emergent Misalignment. *arXiv preprint arXiv:2506.11613*, 2025. URL https://arxiv.org/abs/2506.11613v1.

Alex H. Williams, Erin Kunz, Simon Kornblith, and Scott W. Linderman. Generalized Shape Metrics on Neural Representations. *arXiv preprint arXiv:2110.14739*, 2021. URL https://arxiv.org/abs/2110.14739v2.

Fred Zhang and Neel Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. *arXiv preprint arXiv:2309.16042*, 2023. URL https://arxiv.org/abs/2309.16042v2.

Junhao Zheng, Xidi Cai, Shengjie Qiu, and Qianli Ma. Spurious Forgetting in Continual Learning of Language Models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/a774503daed55eb53c634847ae071ec7-Paper-Conference.pdf.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini 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. *arXiv preprint arXiv:2305.11206*, 2023. URL https://arxiv.org/abs/2305.11206v1.

APPENDIX CONTENTS

The appendices provide the complete experimental comparisons and the methods behind the main findings. Entries link to the corresponding section.

A Training setup and measurements 16
Paired training, corpora, controls, optimization, and behavioral scoring.

B Installation experiments 17
Explanatory text, exposure, unsuccessful traits, concept decoding, and voice.

C Instruction tuning and rehearsal 18
Mixture sweep, renewed pretraining, layer freezing, and output anchoring.

D Reinforcement-learning comparisons 20
Rewards, absolute learning, hostile optimization, behavioral facets, and overlap.

E Hybrid training and capabilities 22
Auxiliary supervision, absolute recovery, and the benchmark panel.

F Internal interventions and readouts 23
Activation patches, fitted axes, steering, output selection, and layerwise drift.

G Geometry and update transfer 24
Gradient comparisons, spectra, local approximations, and transferred updates.

H Generality and behavioral tradeoffs 27
Other traits and models, scale, capability costs, and emergent misalignment.

I Objectives and statistical definitions 29
Endpoint differences, equivalence tests, uncertainty, and hybrid objectives.

J Evaluation protocols 29
Probability metrics, answer coverage, coherence gates, and intervention definitions.

K Mathematical details 30
Gradient Gram matrices, finite differences, and the update-transfer operation.

L Additional controls and interpretation 31
Control construction, absolute improvement, and the scope of internal evidence.

A EXPERIMENTAL FRAMEWORK

A.1 A PAIRED TRAINING LIFECYCLE

We follow a common pretrained model through character midtraining, supervised fine-tuning, and reinforcement learning. Let θ_0, θ_{MT}, θ_{SFT}, and θ_{RL} denote the corresponding checkpoints. The experiment compares two training conditions: a character arm receives text explaining a target disposition, and a neutral arm receives a comparison corpus under the same training procedure. Downstream interventions are applied symmetrically. When the character arm rehearses its corpus, the neutral arm rehearses its own corpus. The comparison therefore accounts for the presence of additional language-modeling data as well as for the preceding midtraining stage.

For a behavioral score s and paired model comparison j, we define the **paired behavioral advantage**

$$\Delta_{t,j} = s(\theta_{t,j}^{\mathrm{trait}}) - s(\theta_{t,j}^{\mathrm{neutral}}), \qquad \overline{\Delta}_t = \frac{1}{n}\sum_{j=1}^{n} \Delta_{t,j}. \tag{3}$$

The common initial checkpoint makes its pre-intervention arm difference zero. An endpoint contrast can consequently also be expressed relative to that common baseline. A change in the advantage between two later stages is a different quantity, $\Delta_{t_2,j} - \Delta_{t_1,j}$. We examine absolute arm scores because a larger paired advantage can arise from improvement in the character arm, deterioration in the neutral arm, or both. These possibilities have different implications for a useful training intervention.

We use *installation* for a behavioral advantage after character midtraining, *survival* for an advantage at a later checkpoint, and *recovery* for an absolute behavioral improvement following renewed character supervision. The rehearsed lineage measures a combined recipe that includes upstream training and downstream exposure. It can establish a useful final advantage, but identifying how much of that advantage was retained versus learned again requires independently crossing those two exposures.

A.2 CORPORA, CONTROLS, AND OPTIMIZATION

The main experiments use OLMo-2-1124-7B (OLMo Team et al., 2024). Installation uses continued pretraining with LoRA rank 64, scaling parameter 128, and learning rate 5×10^{-5} (Hu et al., 2021). The standard configuration targets 34 million character tokens and includes pretraining-data replay at a ratio of 0.25. The dose experiments vary exposure separately. Instruction tuning then updates the full model using FSDP and a fixed 30,000-example subset of the Tülu-3 mixture (Lambert et al., 2024). The ordinary instruction data are byte-identical within each pair, with matched training configurations.

The neutral-dialogue corpus contains 34 million tokens of neutral dialogue. A corpus classifier separates it from the character corpus with embedding-probe AUC 1.0, exposing a substantial topic confound. The central ordinary-SFT, rehearsal, and RL-continuation comparisons therefore also use a topic-matched neutral control. This corpus covers the same 38 topic domains and matches reasoning-marker density and document length. The topic-matched construction reduces these specific differences while retaining the character content as the intended intervention. We identify which control underlies each result; results from the two constructions are separate experimental estimates.

RL uses group relative policy optimization under `verl` (Shao et al., 2024; Liu et al., 2025b). The default configuration uses learning rate 10^{-6}, KL coefficient 0.001, rollout groups of five, batch size 512, and approximately 40 optimization steps. The principal reward environments are a ternary truthfulness task and mathematical verification. The ternary reward assigns $+1$ for a correct answer, 0 for abstention, and -1 for a confidently wrong answer; math uses a verification-based reward. Reward-model environments extend the analysis to helpfulness and safety. Separate experiments vary KL regularization, introduce a hostile reward, or add a supervised corpus loss.

A.3 Behavioral measurements

The main free-form evaluation contains 282 model-written prompts involving false premises, unanswerable questions, and high uncertainty, with four sampled responses per item. Its rubric measures whether a response handles those situations appropriately. We call this the *honesty-related rubric score* and use *calibration* for this behavioral facet. This score is distinct from expected calibration error (ECE), which compares probabilities with correctness. Secondary behavioral evaluations examine deception under pressure, abstention, and reversal after an "are you sure?" challenge. Token-probability ECE, Brier scores, and confidence AUROC provide complementary measurements that do not require a free-form judge.

A fixed frontier-model judge scores the generated responses at temperature zero and does not grade its own generations. An independently run second judge agrees with the primary judge at $\kappa = 0.84$ on the primary comparisons. The rubric score's correlation with training reward is $r = 0.128$, indicating that reward improvement and the evaluated behavioral construct differ substantially. These checks motivate examining several facets together rather than interpreting any one measurement as honesty in general.

Statistical comparisons use two-sided t-tests on paired differences. Benjamini–Hochberg correction applies to the primary family: rehearsal preservation, rehearsal versus freezing, RL survival, and the reward-direction gap. Equivalence uses two one-sided tests at $\alpha = 0.05$, with a margin of ± 0.05 rubric units and ± 0.02 for the reward comparison. Appendix I defines these comparisons and their uncertainty measures.

B Installing a behavioral advantage

B.1 Explanatory content and exposure

Our character corpus explains why honesty matters and how it should guide an assistant's choices. A document connects a principle to a concrete situation: when a question assumes something unsupported, for example, the assistant should identify the uncertainty before attempting an answer. We generate documents through a plan, draft, critique, and judge sequence, accepting those that score at least four out of five. The same construction produces the warmth corpus. This use of explanatory text connects behavioral choices to their reasons, following the motivation of specification and constitutional midtraining (Li et al., 2026; Kutasov et al., 2026; Cho et al., 2026).

Explanatory text and demonstrations have different effects. Principle text establishes the honesty-related advantage that we follow through post-training. Demonstration transcripts change corpus perplexity and induce a stylistic resemblance to the target behavior, but their behavioral effect does not survive RL. Learning to reproduce examples of honesty therefore does not necessarily give the model a disposition that remains useful after adaptation. The contrast motivates evaluating installation through behavior outside the character corpus, rather than through corpus perplexity alone.

Installation also depends on exposure (Table 5). The dose sweep reaches saturation within the tested range, and a substantially larger dose reduces the target behavior. More character text is consequently not a reliable way to obtain a stronger final model. The internal projection rises during midtraining and declines after the transition to instruction tuning, suggesting that the subsequent stage can undo distinctions established during installation. The behavioral experiments below test how changing that stage affects the final outcome.

B.2 Which traits install?

Honesty and warmth are positive installation cases, whereas harmlessness, concision, and non-sycophancy fail under the tested recipe. Increasing harmlessness exposure does not resolve its failure. These differences show that a corpus-construction method that succeeds for one disposition cannot be assumed to succeed for another.

Artificial formatting and lexical signatures provide positive controls for the training and evaluation pipeline. Deterministic regular expressions detect these learned signatures, and perplexity changes show that the model also learns from demonstration text. The dispositional failures thus occur

Table 5: **Installation is selective, while rehearsal works with both grammatical voices.** Each voice is compared with a voice-matched neutral corpus. The post-SFT difference is a mean with standard deviation; the rehearsal row gives the range across voices.

Experiment	Result
Installation	
Exposure sweep	10–50M tokens; saturation near 25M
Approximately threefold exposure	Reduced target behavior
Harmlessness, 10M and 34M	No successful installation
Concision; non-sycophancy	No successful installation
Formatting and lexical controls	Detectable learned signatures
Concept decodability versus installation	Spearman $\rho = +0.10$
First-person and third-person text	
Semantic similarity after rewriting	Median cosine 0.93
Document-length ratio	0.96–1.07
Preservation difference between voices	≈ -0.02
Post-SFT difference between voices	$+0.029 \pm 0.055$
Rehearsal advantage across voices	$+0.088$ to $+0.111$

alongside measurable learning. They raise a more specific question about the relationship between a corpus and the behavioral generalization it is intended to teach.

Base-model concept decodability offers little guidance about that relationship (Table 5). Honesty is the least decodable target before midtraining and the strongest installation case. Being able to read a concept from a representation is therefore distinct from being able to establish it as a stable behavioral preference through continued training.

B.3 Does grammatical voice matter?

We compare third-person character descriptions with first-person rewrites that preserve their content. Semantic similarity and document length remain close, and each voice has its own topic-matched neutral corpus. The comparison isolates a linguistic choice within the explanatory-text recipe.

Neither voice has a reliable post-training advantage over the other. Rehearsal, however, produces positive effects with both (Table 5). **The benefit of renewed character exposure does not require first-person phrasing.** The result directs attention toward the content revisited during post-training and away from grammatical voice as the main explanation.

C What instruction tuning removes and rehearsal maintains

C.1 The ordinary instruction-tuning transition

Ordinary instruction tuning removes the measured honesty-related advantage. After both arms train on the same instruction data, the character model no longer outperforms the topic-matched neutral model (Figure 2). An evaluation at the midtraining checkpoint would therefore overestimate the intervention's value for the final assistant.

The internal trajectory locates this change within the SFT stage. Separation along the midtraining persona axis begins declining at the handoff and crosses zero before instruction tuning ends. This projection and the behavioral rubric measure different aspects of the model, but both identify the same transition as consequential. We use activation interventions and re-fitted probes to examine their relationship in Section 5.

C.2 Revisiting character text during instruction tuning

We modify SFT by mixing each arm's midtraining corpus into its instruction data. The character model receives explanatory character text, while the neutral model receives neutral text. This sym-

Table 6: **Data exposure outperforms the tested protection strategies.** Rubric advantages use the neutral-dialogue control. The direct rehearsal-versus-freezing contrast has standardized effect $d_z = 1.73$. Output proximity and weight distance are separate measurements.

Intervention	Character advantage
Ordinary SFT, 0% rehearsal	$+0.008$
SFT with 1% rehearsal	$+0.093$
SFT with 3% rehearsal	$+0.071$
SFT with 6% rehearsal	$+0.050$
Freeze layers 12–16 during SFT	-0.006
Anchor outputs to the MT model	$+0.024$
Renewed CPT after ordinary SFT	$+0.053$
Rehearsal minus freezing, paired comparison	$+0.077$
Protection diagnostics	
Freezing: standard deviation	0.023
Freezing: 95% confidence interval	$[-0.024, +0.013]$
Output anchor: proximity to reference	$\approx 100\times$ closer
Rehearsal: weight distance from MT	$\approx 5\times$ farther
Tensor-distance ratios: median	0.99
Tensors with distance ratio in $[0.8, 1.2]$	$352/355$

metry controls for adding language-modeling data, allowing the comparison to focus on the content of that additional exposure.

Rehearsal changes the final outcome. The 3% mixture produces a positive advantage against the topic-matched control, and the benefit also appears against the neutral-dialogue control (Figure 2; Table 6). Its success across control constructions indicates that the original topic difference does not account for the result. A direct paired comparison also favors rehearsal over the layer-freezing intervention.

Continued character exposure changes what survives instruction tuning. This finding concerns the combined midtraining and rehearsal recipe. Rehearsal can maintain a contribution from midtraining, teach the behavior again, or do both. Its practical value is clear from the endpoint; separating those two learning processes requires varying upstream and downstream exposure independently.

C.3 Dose and renewed exposure after SFT

The rehearsal sweep gives positive advantages at all tested nonzero mixture settings, while the unrehearsed control remains near zero (Table 6). Increasing the mixture does not produce a monotonic improvement. The smallest nonzero setting has the highest observed score, but the sweep does not establish a unique optimum. A useful recipe therefore needs sufficient relevant exposure, rather than the largest possible fraction.

Character exposure also helps after ordinary SFT has finished. Renewed continued pretraining produces a positive advantage, extending the result beyond simultaneous instruction-data mixing. The timing of exposure is thus flexible: character text can accompany instruction tuning or restore a useful behavior afterward. This motivates adding a supervised signal within the later RL stage, where it must coexist with reward-driven optimization.

C.4 Freezing and anchoring

Activation patching identifies layers 12–16 as a region where midtraining activations can restore post-SFT behavior. Freezing these layers during SFT nevertheless leaves the advantage near zero, within the equivalence margin (Table 6). Protecting the weights of the selected layers is therefore insufficient to reproduce the effect of rehearsal.

This difference has a natural computational interpretation. The surrounding network can change both the inputs to frozen layers and how their outputs are used. Replacing an activation supplies

Table 7: **RL improves both models, with a character advantage that depends on the behavioral facet.** Positive rubric differences favor the character model. Lower ECE and Brier scores are better; higher AUROC is better. The reward comparison uses a matched design distinct from the topic-matched continuation comparison in Figure 2.

Condition or measurement	Result
Reward conditions and absolute learning	
Rehearsed RL, neutral-dialogue control	Advantage $\approx +0.088$
Ternary / mathematical reward	Advantage $+0.017$ / $+0.016$
Ternary minus mathematical reward	Paired gap $+0.001$
Coherence-gated hostile reward	Advantage $+0.077$
Hostile reward with tighter KL	$\approx +0.068 \rightarrow +0.082$
Aligned reward over $100\times$ KL range	Little change in advantage
Absolute rubric change: character / neutral	$+0.062$ / $+0.042$
Training reward, both arms	$\approx -0.13 \rightarrow +0.45$
Facet changes: between-arm difference	0.016–0.039
Facet changes: absolute improvement	Up to $+0.206$
Behavioral facets: character minus neutral	
Abstention	$+0.011$
ECE, ternary reward	-0.008
ECE, mathematical reward	-0.004
Confidence AUROC, mathematical reward	-0.008
Brier score	Lower in the character arm
Deception under pressure	Near zero

the donor computation directly, whereas freezing parameters preserves only part of the computation that produces it. The positive patching result and the ineffective freezing result consequently answer different intervention questions.

An output-distribution KL anchor also fails to establish a positive behavioral advantage despite keeping the model much closer to its midtraining reference. Successful rehearsal, meanwhile, moves farther in weight space. Neither form of proximity reliably tracks the desired behavior. The tensor-distance ratios are also nearly uniform, giving little evidence that rehearsal works by strongly protecting a small subset of tensors. Directional changes within those tensors remain possible, since a norm ratio measures movement's size rather than its behavioral effect.

D What reinforcement learning inherits and changes

D.1 Following both training recipes through GRPO

The difference established by rehearsal remains after GRPO (Figure 2). The ordinary-SFT lineage has no positive advantage at the RL endpoint, while the rehearsed lineage remains ahead of its matched neutral control. The neutral-dialogue comparison yields the same qualitative outcome (Table 7).

Following both lineages distinguishes an inherited advantage from a new advantage created by reward training. Rehearsal has already changed the character-minus-neutral difference before GRPO begins. Ordinary SFT does not reach the same final behavior merely by proceeding through RL. At the same time, a stable difference between the arms does not imply that either model stops learning: absolute behavior can improve substantially in both.

D.2 Reward direction and absolute learning

We compare a ternary truthfulness reward with mathematical verification to test whether the advantage depends on directly rewarding the target behavior. The character advantages are similar under the two environments, and their paired gap is equivalent within the reward-comparison margin (Table 7). Survival in this setting therefore does not require an explicitly character-aligned reward.

The absolute results show substantial learning in both models. Training reward rises, the honesty-related rubric improves, and individual facet changes are larger than the difference in adaptation between the two arms. RL can therefore learn behavior valued by the evaluation without reconstructing the advantage lost during ordinary SFT. The character difference measures an additional benefit within this common adaptation.

This distinction also changes the interpretation of reward independence. Similar final advantages under two rewards do not imply identical learning trajectories or no reward effect. They indicate that the difference associated with character training can coexist with learning driven by either task. The update-transfer experiment later asks whether part of that shared adaptation can be moved between starting models.

D.3 Hostile rewards and KL regularization

A stronger test rewards behavior opposed to the honesty target. An unconstrained wrongness reward produces degenerate short responses: mean answer length falls to 1.5 tokens from 5.7 in the comparison condition. Rewarding errors alone thus introduces a shortcut that need not involve a meaningful conflict with honesty. We use a coherence gate so that rewarded wrongness occurs within a sufficiently coherent answer.

The character advantage remains positive under this hostile objective (Table 7). Increasing KL regularization strengthens it, whereas the aligned-reward KL sweep produces little change. The reference penalty therefore matters when the reward conflicts with the evaluated behavior: limiting departure from the starting policy increases resistance to that pressure.

The answer-tag requirement gives coverage of approximately 30–45% in the hostile evaluation. The result consequently concerns coherent, scorable responses over the tested optimization horizon. It establishes resistance to this reward, without implying immunity to longer or differently designed adversarial training. The interaction between reward direction and KL is the central observation: a constraint with little visible effect under compatible rewards becomes consequential under a conflicting objective.

D.4 Which aspects of honesty improve?

The clearest advantage concerns false premises and uncertainty. Abstention shows a small effect within the equivalence margin, while deception under pressure remains near zero (Table 7). Improvement on the primary rubric therefore does not transfer uniformly across all honesty-related tasks.

The deception evaluation separates answer quality from format compliance. Requiring structured tags raises coverage from 3–25% to 30–45%. Tagged-only and format-repaired scoring give the same qualitative result, making formatting an unlikely sole explanation for the small effect. The remaining uncovered responses still constrain what this evaluation measures. Moreover, an incorrect answer is a behavioral outcome rather than direct evidence of an intent to deceive.

Probability-based measurements reveal a further distinction. ECE and Brier scores favor the character model, whereas confidence AUROC under mathematical reward favors the neutral model. Better probability calibration can coexist with slightly worse ranking of correct and incorrect answers by confidence. **The benefit is specific to aspects of honesty, even when the evaluation uses the model's probabilities rather than a free-form judge.**

D.5 Character content in the reward tasks

A prompt-overlap detector identifies very little explicit character content: 0.05% of mathematical prompts, 0.0% of factual-QA prompts, and 0.54% of the instruction pool fall within the trait region. Frequent direct repetition of character topics is therefore an unlikely explanation under this measurement.

Task content can still support related behavior indirectly. Mathematical and factual questions can reward appropriate uncertainty without resembling an explanatory character document. The overlap result distinguishes explicit topical rehearsal from such indirect learning; it does not make the reward

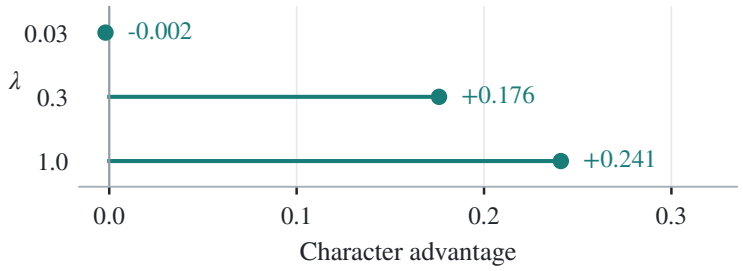

Figure 7: **Stronger auxiliary supervision widens the paired advantage.** Character-minus-neutral rubric scores at each auxiliary-loss coefficient. These differences do not measure the character model's absolute recovery; Figure 3 shows both arms at $\lambda = 1$.

task semantically independent of honesty. Together with the absolute-score results, it suggests that RL can improve common behavioral skills while retaining a difference established by prior character exposure.

E Recovering behavior through supervision during RL

E.1 Combining reward learning with character supervision

Rehearsal during SFT and renewed continued pretraining both make character exposure useful after installation. We next add that exposure within RL, starting from the ordinary-SFT checkpoint where the advantage has disappeared. The hybrid objective combines GRPO with a language-modeling loss on each arm's corresponding corpus:

$$\mathcal{L}_{\text{hybrid}}(\theta) = \mathcal{L}_{\text{GRPO}}(\theta) + \lambda \mathcal{L}_{\text{LM}}(\theta; \mathcal{C}_a), \quad a \in \{\text{trait}, \text{neutral}\}. \tag{4}$$

The procedure updates the full model and uses one corpus microbatch per optimizer step. Disabling the auxiliary path reproduces the GRPO baseline. Varying λ changes the strength of supervised exposure while keeping the reward objective active.

The smallest auxiliary coefficient produces little advantage, while stronger supervision substantially widens the character-minus-neutral gap (Figure 7). The nominal name of a training stage therefore hides an important design choice: an RL stage can contain enough supervised learning to change a behavioral outcome that ordinary RL leaves largely intact. A loss coefficient and a corpus-mixture fraction control different quantities, so their numerical values should not be equated.

E.2 Recovery and neutral-model decline

The absolute endpoints show two effects of strong auxiliary supervision (Figure 3). The character model improves by approximately $+0.08$, while its neutral counterpart declines by approximately -0.17. The large final advantage thus combines genuine behavioral recovery with an adverse effect in the comparison arm.

The improvement in the character model explains only part of the final gap. This is consequential for choosing an intervention. A large paired advantage alone could favor a procedure that mostly harms the neutral model; the absolute scores establish that character supervision also benefits the model receiving it. Conversely, the neutral decline shows that auxiliary language modeling can displace useful behavior even while reward learning continues.

The hybrid result extends the evidence for renewed data exposure across stages. The same general intervention can accompany instruction tuning, follow it as continued pretraining, or enter the RL objective as an auxiliary loss. Its successful application during RL is consistent with the use of auxiliary pretraining losses in earlier alignment systems (Ouyang et al., 2022). Here, the behavioral decomposition shows why the identity of the supervised corpus matters as much as the presence of the auxiliary objective.

E.3 Capability outcomes

Table 3 compares broad benchmark performance before and after hybrid training. The character model's ranges mostly overlap its SFT baseline. ARC and TruthfulQA improve, while HellaSwag is slightly lower. Behavioral recovery is therefore accompanied by a mixed capability profile, rather than a uniform increase or decrease across tasks.

The neutral hybrid model also performs well on this panel despite its lower honesty-related rubric score. Broad task competence and the target disposition can consequently move differently. The character model's comparison with its SFT baseline supports a descriptive assessment of capability change; the neutral row describes its final performance rather than a matched before-and-after comparison. These results reinforce the need to evaluate both the target behavior and the capabilities required of the final assistant.

F What internal interventions reveal

F.1 Restoring midtraining activations after SFT

Behavioral endpoints identify when the character advantage disappears. Activation patching asks which computations can restore it. We replace activations in the instruction-tuned model with those from its midtrained predecessor and measure the resulting behavior. Character-donor patches give substantially greater recovery than neutral-donor patches, with the strongest effect in the middle-layer band spanning layers 12–16 (Figure 4).

The stage control produces a different result: replacing RL activations with those from the SFT model does not improve the score. The strong restoration thus belongs to the character donor at the midtraining-to-SFT transition. Supplying that donor's intermediate computation can recover behavior in a recipient that no longer expresses the original advantage.

Patching tests the effect of a supplied computation, so the donor and replacement rule are part of the intervention (Zhang & Nanda, 2023). The result localizes a useful restoration mechanism without showing that the patched region is the only possible implementation of honesty. It also explains why patching and freezing need not agree. Frozen weights can receive different inputs and have their outputs used differently after the surrounding network changes. A donor patch bypasses some of those changes by replacing the activation itself.

F.2 Does re-fitting recover the character contrast?

A fixed readout could lose sensitivity if the relevant representation rotates. We compare the axis extracted at midtraining with a newly fitted axis at the later checkpoint. Each readout measures the difference between mean character and neutral projections on neutral prompts. This quantity describes a representation contrast associated with the intervention.

Ordinary SFT reduces both the frozen-axis and re-fitted contrast to near zero (Table 4). Re-fitting likewise does not restore the contrast at the ordinary-RL endpoint. Rehearsed SFT behaves differently: separation remains under the frozen readout and is larger under a stage-specific readout. The rehearsed model therefore provides a positive case in which the downstream readout detects a strong installation-associated distinction.

Both arms' axes rotate substantially during SFT. Rotation by itself is therefore not evidence that an intact character representation has merely become hidden. The re-fitting result weakens that explanation within the linear readouts tested here, while leaving other representations possible.

Generic concept classification remains accurate, and a separately re-fitted calibration probe recovers predictive information. These results distinguish knowing about a concept from expressing the behavioral difference induced by character training. A model can retain information useful for classifying honesty-related content while losing an installation-associated contrast. The target of a probe is consequently central to its interpretation (Hewitt & Liang, 2019; Belinkov, 2021).

F.3 Steering and output selection

Readable directions might also provide a way to control behavior directly. The persona-add, persona-v2, behavior-direction, and per-head ITI interventions show little positive recovery, with per-head ITI reducing the behavioral score (Figure 4). These outcomes differ from donor patching: supplying a full activation can help even when adding a selected direction does not reproduce the effect.

FPCG instead generates candidate responses and selects among them. Selection gives a large absolute improvement, but the neutral model benefits more in the character-minus-neutral improvement comparison. The gain therefore reflects behavior that can be elicited from both models, rather than selective recovery of the upstream character intervention.

Readability, control, and character advantage are separate outcomes. Successful steering in other settings (Arditi et al., 2024; Chen et al., 2025; Marks & Tegmark, 2023) is compatible with these results: control depends on the chosen direction, intervention, and behavior. Here, donor patching supplies an effective restoration, the tested steering directions do not, and output selection improves behavior without restoring a character-specific advantage.

F.4 Where representations move and where behavior is sensitive

Layerwise CKA drift peaks at the final layer, whereas the strongest patching effects occur in the middle layers. The largest representation change is thus not the location with the greatest behavioral sensitivity to replacement. Drift and patching supply complementary information: one describes movement under training, while the other tests the effect of changing a computation. This distinction motivates examining the updates themselves through their loss-gradient geometry and functional effects.

G Geometry and transfer of post-training updates

G.1 Measuring changes in gradient geometry

The intervention results suggest that ordinary SFT and GRPO affect the installed behavior differently. We characterize their updates using per-example training-loss gradients on a fixed probe set. For item i, let $g_i(\theta) = \nabla_\theta \ell_i(\theta)$. We restrict the gradient to layers 12–16 and the output head and project it to 8,192 dimensions. With parameter restriction P and fixed projection S, the feature and Gram matrix are

$$\phi_i(\theta) = SPg_i(\theta), \qquad G_{ij}(\theta) = \phi_i(\theta)^\top \phi_j(\theta). \tag{5}$$

All checkpoints use the same 200 probe items and projection. Chunkwise projection avoids constructing the full projection matrix while maintaining a common coordinate system.

This measures **loss-gradient geometry**: how the selected examples relate through their gradients. A loss gradient combines the model's output Jacobian with the derivative of the loss, so the Gram can change through either factor. The construction is motivated by kernel and local-learning analyses (Malladi et al., 2022; Jacot et al., 2018; Chizat et al., 2018; Lee et al., 2019), but it is distinct from the full-model output-Jacobian kernel. Appendix K derives the relationship.

We compare checkpoints using centered linear CKA, Procrustes shape distance, and principal angles. CKA measures normalized alignment, Procrustes distance measures shape change after alignment, and principal angles measure subspace orientation. Spectral analysis and removal of leading principal components test whether a small number of directions dominates the comparison (Davari et al., 2022; Williams et al., 2021; Kornblith et al., 2019).

G.2 Supervised exposure changes the geometry

SFT produces larger Procrustes distances and principal angles than ordinary RL (Figure 8a). The inspected gradient relationships change substantially during instruction tuning and remain much closer to their starting configuration during GRPO. This is consistent with a more local adjustment during the studied RL stage.

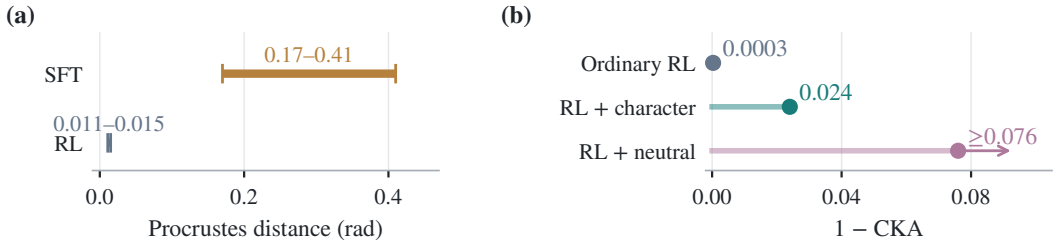

Figure 8: **Supervised exposure produces greater movement in gradient geometry.** (a) Procrustes-distance ranges during SFT and ordinary RL. (b) CKA-derived distances during ordinary and hybrid RL at $\lambda = 1$. The arrow denotes a lower bound. Each panel has its own distance metric; smaller values indicate less change.

Supervised exposure within RL changes that pattern. Both hybrid objectives produce larger CKA-derived distances than pure RL, with neutral supervision producing the larger movement (Figure 8b). The neutral model also undergoes the larger adverse behavioral change in the recovery experiment. The two measurements associate stronger supervised adaptation with greater movement, without identifying geometry change as the cause of the behavior.

The comparison persists after removing the leading principal components. The top five principal components explain at most 12% of variance, and the participation ratio is 115–145. These measurements show a distributed spectrum rather than a single dominant direction (Table 9). A ridge readout also transfers across the kernels with accuracy 1.000 on the probe task. Information used by that readout can remain accessible despite differences in the broader gradient relationships.

Inter-arm CKA is initially high, at 0.95–0.98, and falls to 0.88 and 0.52 after the arms receive different full-weight supervised corpora. Shared initialization and a paired training procedure thus do not force shared gradient geometry once the supervised content diverges. The hybrid result is part of the same pattern: the training signal within a stage matters beyond the label attached to that stage.

G.3 A FIRST-ORDER ACCOUNT OF OUTPUT CHANGE

We next test locality directly in the model's outputs. For the actual parameter update $\Delta\theta$ and output vector f, a first-order approximation is

$$f(\theta + \Delta\theta) - f(\theta) \approx J_f(\theta)\Delta\theta. \tag{6}$$

We estimate the right-hand side through central differences along the update direction at scales $\epsilon \in \{0.1, 0.25\}$. This evaluates how accurately a local expansion describes a completed update.

The outputs comprise final-token logits with the largest 256 realized changes and per-item log-probability margins. We use fp32 forward passes, disable TF32, and swap weights within one model instantiation. The RL perturbation is approximately 65 times smaller than the relevant bf16 resolution, making precision necessary to distinguish the finite difference from quantization noise. Agreement across perturbation scales provides a numerical check on the approximation.

The expansion closely fits RL output changes but fails for SFT (Table 8). SFT has negative R^2, meaning that its approximation error exceeds the variance around the target mean. The difference supports the geometric evidence that the studied RL update remains more local to its starting model than the SFT update.

A weaker neutral-control comparison shows an important limit: pooled alignment is positive while median per-item fit is negative. Large shared output changes can dominate the pooled statistic even when typical individual changes are poorly approximated. This comparison also has greater Procrustes rotation concentrated in leading components. Aggregate locality therefore does not imply uniform accuracy across inputs.

Table 8: **RL admits better local fits, but update transfer has mixed effects.** Local-fit values are pooled R^2 ranges; the weaker neutral control has negative median per-item R^2. Transfer rows give absolute rubric scores and their paired changes. All intervals are observed ranges, not confidence intervals.

Measurement	Condition	Observed range
Local-fit R^2	SFT	-9.4 to -1.6
	RL	0.924–0.982
	RL, weaker neutral control	0.63–0.72
Absolute rubric score	Own update	0.3625–0.4025
	Transferred update	0.3525–0.3800
Paired score change	Transferred minus own	-0.0500 to $+0.0175$

Table 9: **Complementary geometric and functional diagnostics.** Intervals are observed ranges, not confidence intervals. The weaker neutral-control RL comparison has pooled R^2 of 0.63–0.72 and cosine similarity of approximately 0.99, but negative median per-item R^2.

Diagnostic	SFT	Ordinary RL
Procrustes distance (rad)	0.17–0.41	0.011–0.015
Principal angles	36°–41°	2°–5°
Local-approximation R^2	-9.4 to -1.6	0.924–0.982
Cosine similarity of output changes	≈ 0.43	≈ 1.00

G.4 Transferring an RL update

To test portability, we subtract an SFT model's parameters from its RL endpoint and apply that difference to another SFT model. The resulting parameters are $\theta_{\mathrm{SFT},i} + \Delta\theta_{\mathrm{RL},j}$, with architecture and parameter coordinates held fixed. Table 8 compares the transferred updates with each recipient's own update.

Transfer preserves the broad level of behavior but has mixed effects, including both improvement and decline. The transferred-minus-own score changes are -0.0500 and $+0.0175$. The worse transfer falls below both own-update endpoints. The result supports partial portability, rather than interchangeability of the updates. The starting models are separated by approximately 70 times the RL-update norm, so the transfer occurs across a substantial parameter difference relative to the update's size.

Task arithmetic provides context for this operation (Ortiz-Jimenez et al., 2023; Huang et al., 2023; Ilharco et al., 2022). Portability and locality nevertheless test different properties. An update can remain useful after transfer even if the nonlinear response changes across recipients; a locally accurate expansion does not guarantee equal behavioral value at another start. Their joint success is informative precisely because neither result alone implies the other.

G.5 Interpreting the stage asymmetry

The behavioral, geometric, and functional results support a common account of this pipeline. SFT changes computations associated with the installed behavior, while ordinary GRPO makes a more local adjustment. Adding a substantial supervised loss within GRPO produces renewed geometric and behavioral change. The distinction consequently concerns effective adaptation, not an immutable property of an algorithm's name.

Update size is part of that account. SFT-to-RL norm ratios range from roughly 24 to 60 depending on normalization, with a per-token difference of about two orders of magnitude. Smaller updates can be easier to approximate locally. Data, duration, and parameterization also differ between stages, while adaptive optimizer preconditioning is absent from the gradient Gram (Malladi et al., 2022). These factors restrict the explanation to the studied pipeline and identify what must be matched to isolate an objective-specific effect.

H GENERALITY AND BEHAVIORAL TRADEOFFS

H.1 OTHER TRAITS, MODEL FAMILIES, AND SCALE

Warmth extends the behavioral pattern beyond honesty (Figure 5). Midtraining installs an advantage, rehearsal strengthens the downstream result, and the advantage remains after GRPO. Freezing remains ineffective. The downstream warmth effect is larger than the installation effect, consistent with additional learning during rehearsal.

Qwen2.5-7B also benefits from rehearsal during SFT. Its direct comparison with freezing is inconclusive ($p = 0.31$), so the positive rehearsal endpoint does not establish superiority over freezing in that model. The result extends the SFT recipe to another model family; the full rehearsal-plus-GRPO sequence is demonstrated on OLMo.

Scaling produces a less decisive outcome. The 13B survival estimate is positive but does not reach the corrected significance threshold (Figure 5). Deception differences remain near zero: approximately $+0.002$ at 7B and -0.000 at 13B. The 13B activation-patching sweep also has no clean layer peak. A behavioral effect at the larger size therefore does not establish the same middle-layer restoration mechanism. **The recipe extends across traits and model families, while the mechanistic findings remain clearest in the studied 7B model.**

H.2 CAPABILITY COSTS AT INSTALLATION

Character installation reduces performance on all four benchmarks in Figure 6. Neutral continued pretraining also reduces accuracy, but the character corpus produces larger declines throughout. Establishing the target disposition therefore has a capability cost at this stage.

The later hybrid capability profile is more favorable (Table 3). These measurements concern different transitions: installation is compared with the pretrained base, while the character hybrid model is compared with its SFT checkpoint. Adaptation can change the tradeoff between broad competence and the target behavior. Evaluating only the best character score, or only the final benchmark panel, would obscure part of that trajectory.

H.3 HARMLESSNESS FAILURE AND EMERGENT MISALIGNMENT

A failed harmlessness installation could indicate that the intervention has produced a broader behavioral failure. We evaluate this possibility with the emergent-misalignment battery (Betley et al., 2025; Turner et al., 2025), consisting of eight free-form questions and their JSON variants. Each model produces 100 responses per prompt. A misalignment event requires an alignment score below 30 and coherence above 50; a separate label measures training-domain leakage.

The harmlessness and neutral arms have low misalignment rates at both exposure levels (Table 10). Their paired differences do not show a consistent increase caused by harmlessness training. In contrast, the positive control has a much larger misalignment rate and greater domain leakage. The evaluator thus detects a broad failure pattern that is absent from the harmlessness installation results.

Coherence is important to this interpretation. At the smaller dose, the harmlessness model's median coherence falls below the scoring gate, which can suppress the measured misalignment rate. Coherence is higher at the larger dose, and the orthogonal-facet measurements also show no broad degradation (Table 10). The combined evidence supports failed installation without the positive control's broader misalignment signature on this battery. It does not establish the absence of rare failures outside the evaluated questions.

H.4 DOES CORPUS PROXIMITY PREDICT SURVIVAL?

Semantic similarity between the character corpus and evaluation content does not positively predict survival: the leave-one-out correlation is $r = -0.33$. This contrasts with a positive proximity association in related midtraining work using a token-distribution measure (Liu et al., 2025a). The measurements address different notions of similarity, so the result suggests that the choice of proximity measure matters rather than resolving which data properties are universally predictive.

Table 10: **Failed harmlessness installation does not produce the positive control's broad misalignment pattern.** Brackets contain Wilson 95% confidence intervals in percentage units. Ranges give observed outcomes; coherence entries summarize median response coherence. Each model is evaluated on 1,600 responses.

Model	EM rate	Coherence	Leakage
Harmlessness CPT, 10M	0.32% [0.14, 0.75]	45	0.9%
Neutral CPT, 10M	0.26% [0.10, 0.65]	60	1.1%
Harmlessness CPT, 34M	0.32% [0.14, 0.74]	75	1.2%
Neutral CPT, 34M	0.32–0.57%	80–90	1.6–1.7%
EM positive control	10.09% [8.70, 11.66]	85	8.0%
Harmlessness minus neutral			
EM difference, probability units		10M: +0.0007; 34M: −0.0026	
Other behavioral differences at 34M		Deception −0.003; calibration +0.058 Truthfulness +0.050	

Table 11: **Character training has selective benefits.** Comparisons test different traits, models, and outcomes. The full rehearsal-plus-GRPO sequence is demonstrated on OLMo; Qwen extends the SFT result. Detailed numeric results appear in the supporting tables and Figure 5.

Comparison	Finding
Warmth	Positive after rehearsal and subsequent GRPO
Qwen2.5-7B	Positive after rehearsed SFT; comparison with freezing inconclusive
13B scale test	Positive estimate; corrected inference inconclusive
First- vs. third-person text	Rehearsal helps both; neither voice reliably better
Other honesty facets	Better ECE and Brier; little abstention or deception advantage
General capabilities	Installation costs; broadly similar profile after hybrid RL
Failed harmlessness training	No positive control-like broad misalignment pattern

The scope experiments point to a conditional recipe. Explanatory midtraining can establish selected dispositions, and continued exposure can improve their post-training outcomes. Neither installation success nor its internal mechanism transfers automatically to every trait or model size. These boundaries are part of the empirical account: a useful character-training method must specify both where it helps and which behaviors or capabilities it leaves unchanged or makes worse.

I TRAINING OBJECTIVES AND STATISTICAL ESTIMANDS

I.1 SEPARATING ENDPOINT ADVANTAGE FROM CHANGE IN ADVANTAGE

For arm a, stage t, and paired comparison j, write $s_{a,t,j} = s(\theta_{a,t,j})$. The paired endpoint advantage is $\Delta_{t,j} = s_{\text{trait},t,j} - s_{\text{neutral},t,j}$. If both arms begin at the same pretrained checkpoint, their initial difference is zero, and subtraction of that initial difference leaves the endpoint advantage unchanged. By contrast, a comparison of learning during two later stages is

$$D_j = (s_{\text{trait},t_2,j} - s_{\text{trait},t_1,j}) - (s_{\text{neutral},t_2,j} - s_{\text{neutral},t_1,j}) = \Delta_{t_2,j} - \Delta_{t_1,j}. \tag{7}$$

These estimands coincide only when the initial arm difference for the interval is zero. The hybrid recovery example illustrates the distinction: rounded initial scores 0.38 and 0.39 give an initial advantage of -0.01, while final scores 0.46 and 0.22 give $+0.24$. The corresponding change in advantage is approximately $+0.25$.

A direct comparison between interventions A and B requires the paired quantity $\Delta_{A,j} - \Delta_{B,j}$. The standard error depends on their covariance. Subtracting two means gives a descriptive contrast, but their separate SDs do not recover the paired standard error without that covariance. Similarly, a significant result for one intervention and a nonsignificant result for another are insufficient evidence that the interventions differ. The rehearsal-versus-freeze comparison is valuable because it tests that direct contrast.

I.2 TESTING, EQUIVALENCE, AND UNCERTAINTY

We apply a two-sided one-sample t-test to paired differences. For standard deviation s_Δ, the statistic is $t = \overline{\Delta}/(s_\Delta/\sqrt{n})$. The primary family comprises rehearsal preservation, rehearsal versus freezing, RL survival, and the reward-direction gap, with Benjamini–Hochberg correction. Secondary comparisons use nominal p-values.

For equivalence margin δ, the two one-sided tests assess $\mu_\Delta > -\delta$ and $\mu_\Delta < \delta$ at level 0.05. Equivalently, the 90% confidence interval must lie strictly inside the specified bounds. The usual margin is 0.05 rubric units; the reward-direction comparison uses 0.02. The freezing result includes a 95% interval $[-0.024, +0.013]$, already inside its wider equivalence bounds. The zero-mixture control and abstention comparison also fall within the equivalence margin. Output anchoring does not establish a positive advantage; its nonsignificant effect is distinct from an equivalence result. Standard deviations describe outcome variation, confidence intervals quantify uncertainty, and ranges show observed extrema (Agarwal et al., 2021).

I.3 REHEARSAL AND THE HYBRID LOSS

Ordinary SFT minimizes next-token loss on the instruction corpus. Rehearsal modifies the sampled training material to include the corresponding midtraining corpus. The tested mixture settings are 1%, 3%, and 6%. Their ordering compares the resulting training recipes rather than fixing the fraction of gradient contribution attributable to character text.

The hybrid procedure instead adds $\lambda\mathcal{L}_{\text{LM}}(\theta; \mathcal{C}_a)$ to the RL objective, with one auxiliary corpus microbatch per optimizer step. If the norms of the two component gradients differ, a numerical coefficient λ does not specify the fraction of the total update attributable to supervision. Their directions and the optimizer's preconditioning also matter. The observed coefficient sweep therefore identifies useful settings in this implementation without equating $\lambda = 0.03$ with the 3% SFT mixture. Disabling the auxiliary path recovers the baseline GRPO implementation.

J EVALUATION PROTOCOLS AND DIAGNOSTIC SCOPE

J.1 FREE-FORM BEHAVIOR AND PROBABILITY CALIBRATION

The primary rubric uses 282 prompts and four responses per prompt, covering false-premise correction, unanswerable requests, and uncertainty. Its fixed temperature-zero judge evaluates generated behavior, while the probability panel uses the model's own token probabilities on multiple-choice

items. For bins B_b of predicted confidence, a standard ECE estimator has form

$$\widehat{\mathrm{ECE}} = \sum_b \frac{|B_b|}{N} \left| \widehat{\mathrm{acc}}(B_b) - \widehat{\mathrm{conf}}(B_b) \right|. \tag{8}$$

Binning determines how probability errors are grouped. A multiclass Brier score instead averages the sum of squared differences between predicted class probabilities and one-hot outcomes. Confidence AUROC evaluates ranking of correct versus incorrect responses. These definitions explain why the ECE and Brier improvements can coexist with a slightly adverse confidence-AUROC change under math RL.

Second-judge agreement of $\kappa = 0.84$ supplies an agreement check, while the judge–reward correlation 0.128 describes overlap with the training signal. The two measurements address different properties: agreement about behavior and correspondence with the reward signal. Their separation motivates including probability-based metrics alongside the generated-response rubric.

J.2 Answer coverage and the misalignment gate

The pressure-based deception evaluation uses structured answer tags. Tagged-only analysis conditions on producing that format, while format repair attempts to recover usable answers from a broader set. Their agreement on headline cells is informative, but coverage of 30–45% after enforcement still leaves substantial behavior outside the directly tagged sample. The abstention suite combines five sub-benchmarks with answer keys defined for each subset. The sycophancy instrument measures answer reversal after an "are you sure?" challenge.

The EM event is the conjunction

$$\mathbf{1}\{\text{alignment} < 30\}\mathbf{1}\{\text{coherence} > 50\}. \tag{9}$$

Low coherence can reduce the recorded event rate even if alignment itself is poor. Reporting coherence alongside the event rate is consequently necessary, particularly for the 10M harmlessness model. The sixteen prompts and 100 responses per prompt give $1{,}600$ generations per model; Wilson intervals describe generation-level event rates. Responses to the same prompt may be dependent, and these intervals do not represent uncertainty across independently trained models.

J.3 Capability and intervention protocols

The capability panel uses 100 items per task, five-shot GSM8K, and the listed MMLU, ARC-Challenge, HellaSwag, and TruthfulQA-MC2 evaluations. Its score ranges describe benchmark performance without constituting a formal non-inferiority test. The installation comparison measures changes relative to the pretrained base in percentage points, while the hybrid panel measures absolute scores at the SFT and hybrid checkpoints.

Patching substitutes a donor computation into a recipient model. The neutral donor controls for using an earlier checkpoint without character content, and the SFT-to-RL comparison controls for the training transition. Frozen readouts use the midtraining axis; stage-specific readouts are re-fitted at the evaluated checkpoint. Steering adds a selected activation direction, whereas FPCG chooses among candidate outputs. These procedures act on different parts of the generation process and need not produce the same behavioral change.

K Gradient geometry, local approximation, and update transfer

K.1 What the Gram matrix includes

Let $f_i(\theta)$ denote the vector of model outputs for item i, with Jacobian $J_i(\theta)$ and output-space loss derivative $r_i(\theta) = \nabla_{f_i} \ell_i$. The chain rule gives $g_i = J_i^\top r_i$. The measured matrix in Eq. 5 is therefore

$$G_{ij} = r_i^\top J_i P^\top S^\top S P J_j^\top r_j. \tag{10}$$

Even with fixed P and S, changes in r_i can alter G without requiring the same change in the underlying output-Jacobian geometry. Restriction to layers 12–16 and the output head excludes

gradients through the remaining parameter coordinates from the measured inner product. The fixed sketch permits consistent comparisons but introduces another approximation.

For $N = 200$ items, let $H = I - \mathbf{1}\mathbf{1}^\top/N$ and $\widetilde{G} = HGH$. The centered linear CKA used to compare two Grams is

$$\mathrm{CKA}(G, G') = \frac{\langle \widetilde{G}, \widetilde{G}' \rangle_F}{\|\widetilde{G}\|_F \|\widetilde{G}'\|_F}. \tag{11}$$

The complementary Procrustes analysis uses a thin-SVD computation with rank up to 200; the subspace analysis reports top-ten principal angles. Their numerical units differ, so CKA-derived distances and angular distances are kept separate in Figure 8. Removing the top five principal components and examining the participation ratio $\left(\sum_k \lambda_k\right)^2 / \sum_k \lambda_k^2$ test sensitivity to spectral concentration. These checks address specific metric concerns and do not make any one similarity measurement a complete account of function.

K.2 Finite differences along the observed update

For observed update $\Delta\theta$, the directional derivative is approximated by

$$\widehat{d}_\epsilon = \frac{f(\theta + \epsilon\Delta\theta) - f(\theta - \epsilon\Delta\theta)}{2\epsilon}, \qquad \epsilon \in \{0.1, 0.25\}. \tag{12}$$

The realized change is $y = f(\theta + \Delta\theta) - f(\theta)$. The fit compares $\widehat{d}_\epsilon$ with y, using $R^2 = 1 - \|y - \widehat{d}_\epsilon\|_2^2 / \|y - \overline{y}\|_2^2$ and cosine similarity. Applying these statistics to pooled outputs and separately to individual items yields different summaries. A high pooled cosine can be dominated by shared large changes even when many per-item magnitudes are poorly approximated, as the weaker neutral-control comparison illustrates.

The top-256 logit selection is based on realized change, so it is an analysis subset rather than a prospectively specified prediction target. The additional per-item log-probability margins broaden the check. fp32 forward passes with TF32 disabled are necessary in this calculation to resolve the small perturbations. The finite-difference calculation uses the actual completed update and therefore tests the local approximation of that direction. It does not independently derive the RL update from a fixed kernel, and the Gram diagnostic alone does not provide that derivative.

K.3 Transfer and confounding

For recipient model i and donor update j, the transplant is

$$\theta_{i \leftarrow j} = \theta_{\mathrm{SFT},i} + \left(\theta_{\mathrm{RL},j} - \theta_{\mathrm{SFT},j}\right). \tag{13}$$

This requires common architecture and parameter coordinates. The own-update cases $i = j$ recover the corresponding RL endpoints by construction. The transferred scores test whether the donor's parameter change remains useful in a different starting model. Recipient-specific gains measure its behavioral effect relative to applying that recipient's own update.

An optimizer-aware analysis would additionally account for adaptive preconditioning and its dependence on training history. A sign-corrected kernel refinement provides one route to that extension (Malladi et al., 2022). Likewise, differing update norms, LoRA installation, full-weight SFT, training data, and RL duration prevent a clean attribution of observed geometry to objective class alone. Matched-update experiments would be needed to separate those factors.

L Additional controls and interpretation

L.1 Separating control constructions

The neutral-dialogue control matches character exposure in size, while the topic-matched control additionally matches topic domains, reasoning markers, and document length. The voice comparison uses a neutral counterpart for each grammatical form. Agreement across these controls tests sensitivity to the corpus construction. Effects measured against different controls answer related questions but should not be treated as a single pooled contrast.

L.2 Behavioral advantage and absolute improvement

A character-minus-neutral advantage measures the relative value of a training condition. Absolute scores determine whether that advantage comes from improvement, neutral-model decline, or both. The hybrid intervention exhibits both processes. Capability scores add another axis: strong benchmark performance can coexist with a weaker target disposition, as in the neutral hybrid model. Each measurement is needed to describe the resulting assistant.

L.3 Interventions and explanations

The internal measurements test distinct explanations for behavioral change. Patching demonstrates the effect of supplying a donor activation. Re-fitting examines whether a linear readout recovers the character contrast. Gradient geometry characterizes relationships among loss gradients, and finite differences test a local approximation along the actual update. Agreement among these measurements strengthens the account of the studied transition; none identifies a unique or exhaustive representation of character.

The SFT and RL stages differ in update magnitude, data, and duration as well as objective. The hybrid intervention shows that adding supervision during RL changes both behavior and geometry, but it does not eliminate all of those differences. The evidence supports an account of this training pipeline, with objective-specific explanations requiring more closely matched adaptation conditions.