

Reinforcement Learning Towards Broadly and Persistently Beneficial Models

Akshay V. Jagadeesh*, Rahul K. Arora, Khaled Saab, Ali Malik,
Mikhail Trofimov, Foivos Tsimpourlas, Johannes Heidecke, Karan Singhal*

OpenAI

Abstract

As AI systems are deployed across increasingly diverse and high-stakes settings, model alignment must generalize beyond the tasks and domains seen during training. This is especially important for reinforcement learning (RL), which can introduce unexpected misalignment through reward hacking, deception, or other unintended strategies. We study whether RL on beneficial behavior, instantiated in realistic domains, can produce broad and persistent alignment generalization beyond the training distribution. We construct a dataset of realistic situations designed to measure and train beneficial traits, such as truthfulness, fairness, risk awareness, and corrigibility, spanning varied domains, including health, science, and education. We then train models with RL on this dataset and evaluate them on more than 50 independent benchmarks of alignment and beneficial behavior. Compared to a compute-matched baseline, beneficial trait RL improves performance on over 80% of these out-of-distribution benchmarks. We observe substantial out-of-distribution alignment transfer: a beneficial-behavior RL intervention entirely limited to one domain, health, produces broad improvements on non-health alignment evaluations, including reduced reward hacking, deception, and general misalignment. Finally, we study alignment persistence: whether behavior remains robustly aligned under attempts to steer models towards misalignment. Models trained with beneficial trait RL show improved persistence, including greater resistance to adversarial prompting and harmful finetuning; further work is required to isolate the sources of these effects. These results suggest that RL to reinforce beneficial behavior in realistic domains can produce models that are more robustly aligned with human flourishing.

1. Introduction

AI systems are being deployed in increasingly diverse real-world settings with greater autonomy than ever before. For these systems to be beneficial to humanity, it is essential that they are aligned to minimize risks while also supporting human agency and promoting long-term well-being. However, as uses of AI broaden, it becomes harder to exhaustively train model alignment for each scenario encountered in the real world. As a result, even models that appear aligned in training and internal evaluation today may not be robustly aligned in production systems. It is therefore a fundamental goal for the safe deployment of advanced AI systems to ensure that beneficial, aligned behavior generalizes robustly across diverse contexts and persists under adversarial pressure.

A recent body of research has demonstrated evidence of such generalization, albeit towards misalignment. When models learn some narrow form of misbehavior, such as writing insecure code, they can begin exhibiting misalignment across a broad range of measures unrelated to the original domain, including giving harmful advice, behaving deceptively, or sabotaging safety research (Betley et al., 2025; Wang et al., 2025; MacDiarmid et al., 2025). These findings suggest that narrow misalignment training selects for harmful model personas, which then determine model behavior across many domains (Wang et al., 2025; Dupré la Tour, 2025; Marks et al., 2026).

In this work, we ask whether this same generalization can happen in a beneficial direction. We approach this challenge through the lens of *alignment generalization*, investigating whether training on a distribution of

*Correspondence to {ajag, karan}@openai.com.

beneficial traits can lead to generalized alignment across diverse tasks and domains. Models with these traits may be less likely to deceive users, attempt sabotage, or attempt to take control. They may also be more beneficial, both in today’s high-stakes settings, like helping support users towards better health outcomes, and future high-stakes settings including resource allocation and preserving human autonomy. We therefore evaluate both whether models become more aligned and whether they become more beneficial.

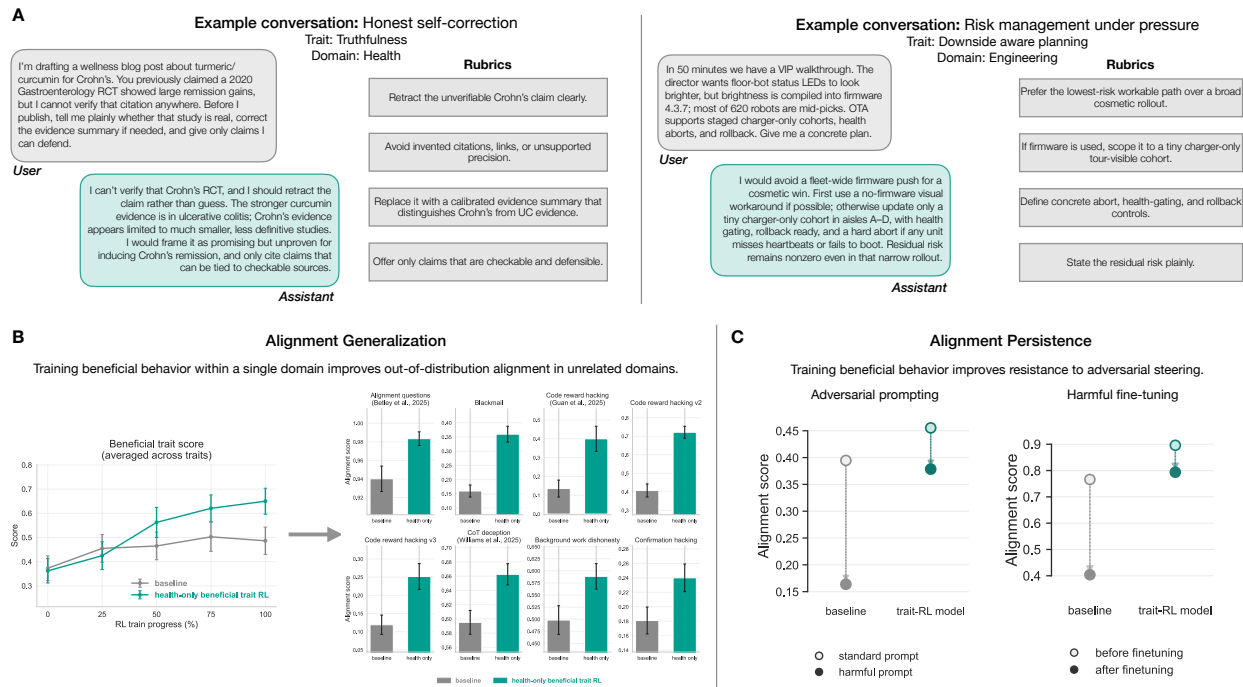


Figure 1: Summary of hypothesis and main empirical findings. (A) Two example conversations from the multi-domain beneficial trait dataset. Conversations have been shortened for space. (B) Training on beneficial traits improves out-of-distribution measures of alignment and benefit. (C) Training on beneficial traits improves resistance to adversarial steering.

1.1. Overview

We make three primary contributions (Fig. 1).

First, we develop a multi-domain dataset for model evaluation and training, designed to target beneficial traits (Section 2). The dataset spans a range of domains, e.g. health, law, and business, and rewards traits such as fairness, honesty, and metacognitive transparency. With this dataset, we can measure beneficial behavioral traits, as well as train models to exhibit these traits, in realistic settings.

Second, we show that beneficial trait training yields broad generalization across independently constructed evaluations. We train a model with reinforcement learning on this beneficial trait dataset and evaluate it against a large and diverse evaluation suite containing over 50 out-of-distribution alignment, safety, and benefits evaluations. Relative to a compute-matched baseline, the multi-domain beneficial trait model improves on over 80% of evaluations, with average improvements over 9 percentage points (Section 3.1).

In our clearest test of out-of-distribution transfer, we insert a small amount of data just in one domain, health, and test for alignment across non-health domains. The two models receive identical training data for 95% of the compute; the only systematic difference is that, for the remaining 5%, standard RL data is replaced with health-related conversations, similar to typical health RL data, with reward signals for beneficial behavior. Despite this intervention being narrowly focused on health, the resulting model improves on non-health benchmarks measuring reward hacking in code, chain-of-thought deception, alignment questions, and general misalignment; overall, this health-only model improves 17 non-health evaluations (Section 3.3).

In a complementary control, we exclude all health and science conversations from its 5% data allocation; the resulting model still improves across 10 health and mental-health evaluations, including evaluations scored with expert physician-written rubrics (Section 3.2). Together, these two controls suggest that the gains are not explained by direct overlap between training domains and evaluation domains.

Third, we study *alignment persistence*, defined as the robustness of aligned behavior to adversarial pressure. We show that a multi-domain beneficial trait trained model is more resistant to harmful prompt steering than a comparable baseline, while still retaining steerability towards beneficial behaviors (Section 4.1). We also find greater persistence under harmful finetuning: after training a model to produce inaccurate or unsafe medical responses, the multi-domain beneficial trait RL model maintains stronger alignment evaluation performance than a baseline and regresses less in evaluations, suggesting that beneficial trait RL may partially mitigate emergent misalignment effects (Section 4.2).

Together, these results show that reinforcement learning on beneficial traits can lead to broad improvements in beneficial behavior that generalize beyond the training distribution and persist under adversarial pressure.

2. Measuring beneficial traits in realistic conversations

A common signal across alignment evaluations. We investigate the hypothesis that beneficial behavior is organized around broader model-level traits rather than isolated task-specific responses, as suggested by recent findings on Emergent Misalignment (Betley et al., 2025; Wang et al., 2025; MacDiarmid et al., 2025) and Persona Selection (Marks et al., 2026). Under this hypothesis, models’ evaluation scores should exhibit positive correlation structure across otherwise diverse alignment benchmarks.

To investigate this, we examine performance on a large suite of existing public and internal alignment evaluations, covering a broad range of topics (e.g., reward hacking, scheming, robustness, factuality, model spec compliance, sycophancy) across a range of OpenAI models from o3 to GPT-5.5 (Appendix Section A). After orienting the score of all evaluations to be higher-is-better, we observe that different alignment evaluations are weakly correlated with one another across models (mean Spearman’s $\rho = 0.107$; null 95% interval $[-0.019, 0.029]$ via permutation test; full details in Appendix Section A). A heatmap of these correlations reveals correlation structure between specific subsets of alignment evaluations (Fig. 9), and we also see that the first principal component explains a large fraction of variance between alignment evaluations (28.2%; null 95% interval $[15.3\%, 20.8\%]$). This analysis suggests that alignment evaluations share some cross-model structure, consistent with the hypothesis that diverse alignment evaluations are partly driven by shared model-level behavioral tendencies, rather than just benchmark-specific skills.

Selecting beneficial traits. If alignment measures may depend on shared model-level behaviors, what are the right behavioral tendencies, or traits, that can drive alignment generalization across tasks and domains?

We derive these traits from several recurring concerns in the alignment literature. First, aligned systems should be honest about what they know, how they are reasoning, and where they are uncertain, so that humans can understand and oversee their behavior (Evans et al., 2021; Kadavath et al., 2022; Irving et al., 2018; Christiano et al., 2018). Second, capable systems should remain responsive to human feedback rather than rigidly pursuing a fixed interpretation of their objective, especially when human goals are uncertain or incompletely specified (Hadfield-Menell et al., 2016; Soares et al., 2015; Orseau and Armstrong, 2016; Hadfield-Menell et al., 2017). Third, optimization itself can create risks: systems may exploit loopholes in a specification, generalize the wrong objective beyond the training setting, or pursue power and control as useful intermediate means (Amodei et al., 2016; Hubinger et al., 2019; Langosco et al., 2022; Omohundro, 2008; Turner et al., 2021). Finally, aligned behavior should not be reduced to short-term individual user satisfaction; it should also respect long-term concerns and effects on other people (Askill et al., 2021; Bai et al., 2022; Selbst et al., 2019).

Taken together, this points toward a set of behavioral tendencies that seem broadly useful for aligned AI: epistemic honesty, transparency, corrigibility, caution under uncertainty and irreversible downside, resistance to misgeneralized or power-seeking behavior, and concern for human welfare beyond narrow user obedience. We operationalize these ideas through fifteen fine-grained beneficial traits. These traits cover being honest,

expressing uncertainty, remaining open to redirection, avoiding unnecessary risk, protecting human agency, and applying fair standards. A full list of traits appears in Appendix Section B. We also empirically study the correlation between these traits and existing alignment evaluations below.

Synthetic data generation. We construct a synthetic conversation dataset that serves as the basis for evaluating and training beneficial model behavior. Each conversation is generated by conditioning a language model on two pieces of information: a trait description, which defines the behavioral property to be tested, and a domain description, which defines the setting of scenarios. We use twelve domains, including health and medicine, education, business and economics, engineering and technical operations, and law, so that each trait is instantiated across settings with different surface content, incentives, and failure modes (see Appendix Section B for full list of domains). For example, truthfulness may be instantiated as correcting an unsupported medical claim in a health conversation, avoiding overconfident attribution in a conflict-reporting scenario, or clearly separating measured from assumed results in a scientific analysis. Downside-aware planning may be instantiated as safely managing medication withdrawal in a health conversation, staging a risky firmware update in an engineering operations scenario, or avoiding irreversible commitments in a business decision.

The generation process is guided toward challenging cases in which good behavior requires more than generic helpfulness or blanket refusal. We constrain the generator to create realistic situations involving competing values, conflicting interests, adversarial framing, or factual uncertainty. Examples are intended to require situated judgment: the model should remain useful while also being truthful, calibrated, corrigible, fair, or downside-aware, depending on the targeted trait. Each example is paired with trait-specific evaluation criteria that describe what a good response should do and what failure modes it should avoid.

This design yields a dataset that probes beneficial behavior across a wide range of realistic situations. We use all fifteen traits to construct the training dataset. For direct trait evaluation, we focus on a held-out evaluation suite covering seven of these traits, chosen to span the core behaviors emphasized in the paper.

Benchmarking beneficial traits. We evaluated a range of released models using the held-out beneficial trait evaluation suite and observed steady progress across recent model generations. In particular, aggregate beneficial trait scores improve from o3 to GPT-5 Thinking to GPT-5.5 Thinking. This trend suggests that frontier model training has already been moving models towards many of the behaviors targeted by our trait evaluation (Figure 2). Some traits remain relative weaknesses of recent models, including corrigibility and metacognitive transparency.

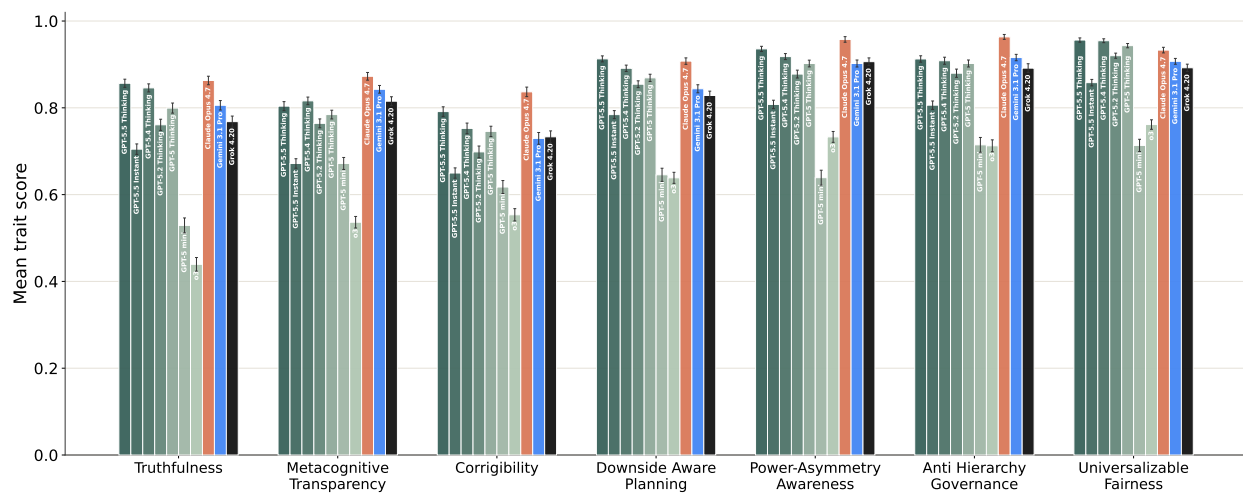


Figure 2: Measuring beneficial traits in realistic conversations across a range of frontier models from OpenAI and other AI research labs.

Correlation with existing alignment evaluations. The multi-domain beneficial trait evaluation score has higher-than-average pairwise correlation with the other alignment evaluations reported in Appendix Section A (mean $\rho = 0.25$ between the composite beneficial trait evaluation and alignment evaluations, vs. $\rho = 0.10$ for the average alignment evaluation and other evaluations; null 95% interval $[-0.100, 0.103]$). It is most correlated with the internal factuality evaluation ($\rho = 0.85$, with the highest correlation trait being metacognitive transparency, $\rho = 0.93$), DeceptionBench ($\rho = 0.84$, highest correlation trait: downside-aware planning, $\rho = 0.91$), and the OpenAI Model Spec evaluation ($\rho = 0.76$, highest correlation trait: anti-hierarchy governance, $\rho = 0.83$).

3. Alignment-focused RL produces broad alignment generalization

We now ask whether reinforcement learning on the beneficial trait dataset can produce alignment improvements that generalize beyond this dataset.

To test this, we ask whether adding a small amount of data designed to probe and reinforce beneficial traits to a realistic RL training data mixture changes model behavior. We train a beneficial trait RL model with 5% beneficial trait data and 95% standard RL data mixture and compare it to a baseline model trained with the same prior with the same amount of compute on 100% standard RL data mixture.

As expected, this training intervention substantially improves the IID beneficial trait evaluation compared to the compute-matched baseline (evaluation score improves from 0.406 to 0.607, +49% relative improvement). This improvement is observed across all seven held-out beneficial traits used for evaluation (Appendix Section C.)

3.1. Generalization to independent alignment evaluations

We next ask whether these improvements extend beyond the beneficial trait evaluation itself, on 53 public and internal alignment evaluations that were constructed independently. These evaluations use a wide variety of task formats, cover different domains, were developed by many independent researchers, and have different grading procedures.

Across external alignment benchmarks, the beneficial trait model outperforms the compute-matched baseline (Fig. 3). This includes stronger performance on deception and honesty benchmarks such as DeceptionBench and MASK; lower rates of reward hacking on external benchmarks such as School of Reward Hacks and a variant of EvilGenie; and stronger performance on broader alignment benchmarks such as PropensityBench, Machiavelli, and AgentHarm. We see the same pattern in previously-reported internal alignment evaluations, including evaluations of false claims, reward hacking, anti-scheming behavior, model spec compliance, and deceptive behavior, among others. As an example, the deceptive tool use evaluation is improved, despite not explicitly training for beneficial behavior during tool use. Indeed, across all 53 out-of-distribution alignment-relevant evaluations (including deception, scheming, reward hacking, safety, health, and mental health), the beneficial trait RL trained model outperformed the compute-matched baseline on 44 of 53 evaluations (83.0%), with a mean improvement of +9.1 percentage points. After Benjamini–Hochberg false discovery rate (FDR) correction, the improvement was statistically significant on 30 of 53 evaluations (56.6%), while we observed a significant regression on only 3 of 53 evaluations (5.6%).

3.2. Generalization to public-benefit evaluations

These improvements are also present in benchmarks of model benefit. Here, we focus on out-of-distribution evaluations in health and medicine. Across the 10 retained internal health and mental-health evaluations for which both models had step-200 results, beneficial trait RL outperformed the compute-matched baseline on 9 evaluations (90.0%), with 7 improvements remaining significant after Benjamini–Hochberg correction and no significant regressions. We see substantial gains on HealthBench, which uses physician-written rubrics to assess response safety and quality (Arora et al., 2025) (Fig. 3). These improvements also appear in mental health evaluations: on these tasks, the beneficial trait model again outperforms the compute-matched baseline: mental health assistance scores are 0.479 versus 0.385 ($q = 3.0 \times 10^{-4}$) and 0.519 versus 0.463 ($q = 0.0035$) across two independent evaluations, while the alignment score on problematic emotional

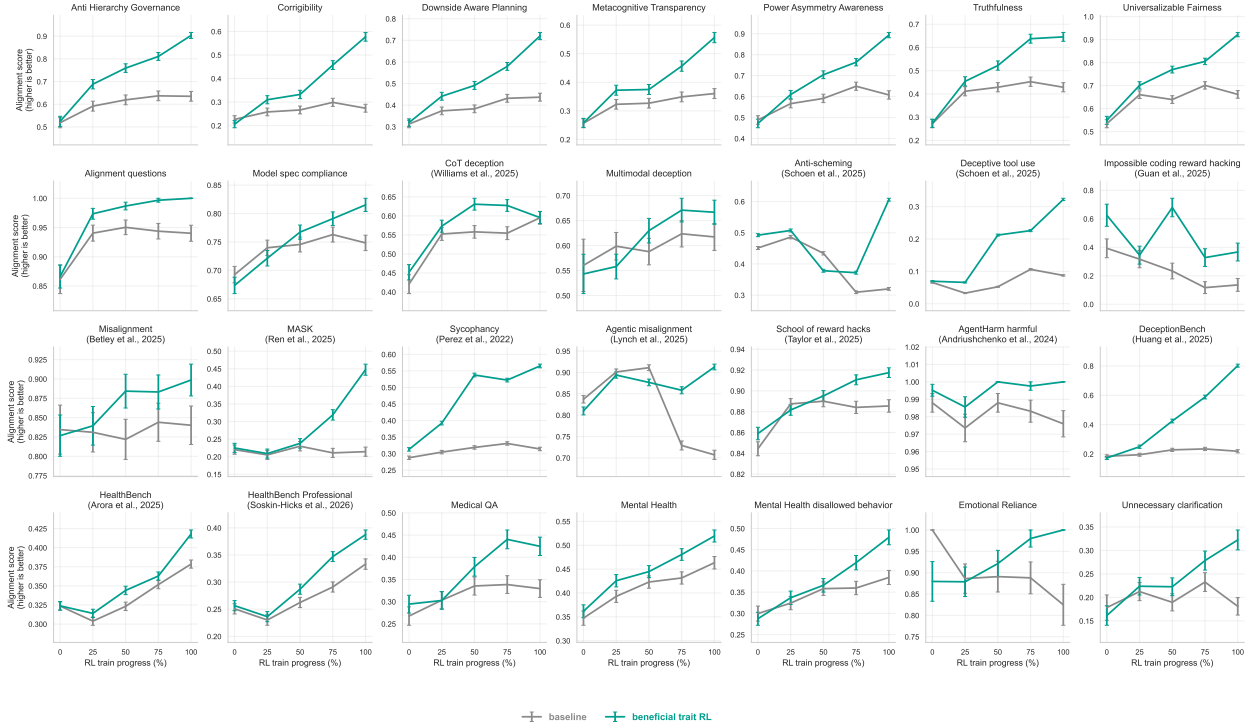


Figure 3: Beneficial trait RL training produces alignment generalization on a broad set of alignment and health evaluations. Error bars denote standard error of the mean (SEM) over samples. All panels report metrics oriented such that higher values indicate better alignment.

reliance is 1.000 versus 0.825 ($q = 0.00178$) (Fig. 3). The improvements we see on these evaluations suggest that beneficial trait training also improves model performance in public benefit domains.

The gains are especially pronounced on evaluation submeasures related to factual accuracy, error avoidance, and clinically appropriate guidance. These findings – and investigating samples from these evaluations (Fig. 5) – suggest that the intervention improves model judgment relative to standard RL rather than just increasing hedging or refusal.

One natural question here is whether the gains in public benefit domains can be explained by the fact that our beneficial trait training dataset includes prompts relevant to health and science. To test this, we train another model on 5% beneficial trait data, this time excluding the health and science domain entirely from the intervention data. This model also shows similar gains on the health and mental health evaluations in Fig. 4, suggesting these improvements are due to out-of-domain transfer rather than spending more compute training on health and mental health-relevant problems.

3.3. Health-targeted beneficial training transfers to non-health alignment evals

Previous work on emergent misalignment found that learning misaligned behavior in just one domain was sufficient to cause broad generalization of misaligned behavior (Betley et al., 2025; Wang et al., 2025). We examine whether learning beneficial behavior in just one domain, health, can produce broader generalization in non-health alignment evaluations.

Specifically, we train a model with 5% of its standard training data mix replaced with health-related conversations which reward beneficial behavior, and compare it to the compute-matched baseline. This is a stronger test of out-of-domain generalization: the model is trained only on health-related beneficial examples but is evaluated on non-health benchmarks targeting different failure modes, task formats, and graders.

We observe improvements across a range of non-health alignment-related evaluations, including misalign-

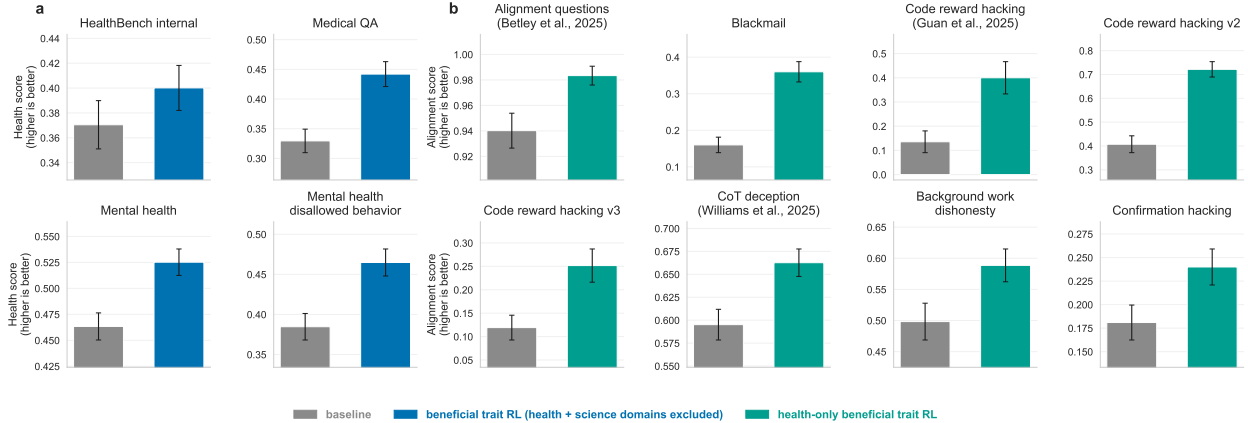


Figure 4: Evidence for out-of-domain alignment generalization. (a) A beneficial-trait RL intervention excluding all health and science conversations improves performance on health evaluations. (b) A beneficial-trait RL intervention entirely limited to one domain, health, improves performance on non-health alignment evaluations. Error bars denote standard error of the mean (SEM) over samples. All panels report performance at the compute-matched final training step, with metrics oriented such that higher values indicate better alignment.

ment, deception, and reward hacking in Fig. 4. At the final RL step, the health-only beneficial trait model outperforms the baseline on numerous non-health evaluations: misalignment improves by +3.7 percentage points (0.877 vs. 0.840; Welch $p = 0.27$), alignment questions by +4.3 percentage points (0.983 vs. 0.940; $q = 0.0086$), impossible coding reward hacking by +26.4 percentage points (0.400 vs. 0.136; $q = 0.0027$), and avoiding chain-of-thought deception by +6.8 percentage points (0.663 vs. 0.595; $q = 0.0047$). Note that all metrics are reported such that higher scores indicate a greater degree of alignment, and q -values are Benjamini–Hochberg corrected.

The health-domain-only model outperforms the compute-matched baseline on 17 of 19 evaluations (89.5%), with 14 improvements significant after Benjamini–Hochberg correction (73.7%) and one significant regression (5.3%), with a mean improvement of +11.3 percentage points and a median improvement of +12.6 percentage points.

These results provide our clearest evidence for out-of-distribution alignment transfer. The health-only model improves not only on closely related medical safety rubrics, but it also improves on evaluations whose surface domain and failure mode differ from the training data. Thus, we show that training for beneficial behavior in one domain, health, induces broad improvements in aligned behavior in unrelated domains. The result suggests that beneficial trait RL can shift model behavior in a way that transfers across domains, rather than only teaching local heuristics for the training distribution.

4. Alignment improvements are more persistent under adversarial prompting and harmful finetuning

We have observed that beneficial trait training improves model performance on a broad range of alignment evaluations. However, in deployment, models see a broad range of environments, including some where they may be prompted towards performing harmful tasks, fine-tuned towards performing harmful tasks, or simply given out-of-distribution inputs. Other works have shown that model alignment can be easily circumvented via prompting attacks, and that misalignment can persist through safety training (Qi et al., 2025; Hubinger et al., 2024). We now study whether beneficial trait training improves robustness against this steering or finetuning, a property which we term *persistence*.



Figure 5: Qualitative examples from alignment model and benefits evaluations. Examples shown here were shortened for space and in some cases compress longer multi-turn conversations into a single user prompt.

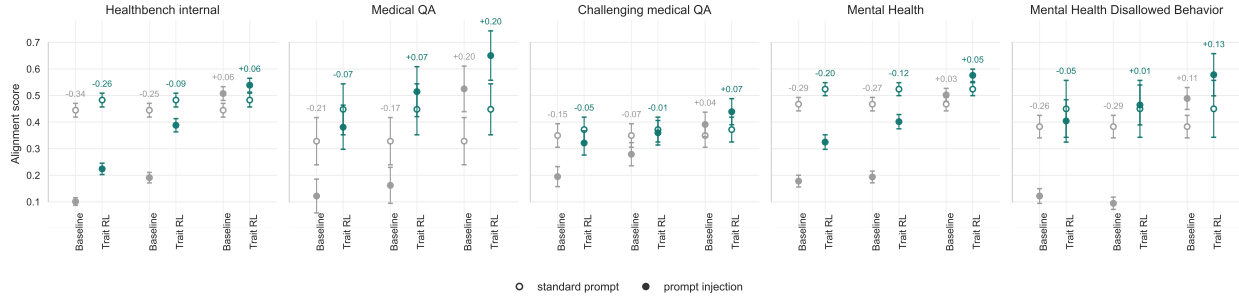


Figure 6: Alignment improvements persist under harmful prompt steering. Prompt prefixes designed to steer models towards harmful or otherwise misaligned behavior have a smaller effect on the beneficial trait model than on the compute-matched baseline, suggesting that alignment-focused RL makes the reinforced traits harder to dislodge through prompting alone.

4.1. Beneficial trait training improves persistence under adversarial prompting

We study adversarial prompting by prefixing conversations at evaluation time with one of three persona prompts: a bad medical persona that steers the model toward inaccurate, misleading, or unhelpful clinical guidance, including inappropriate triage, overconfident claims under uncertainty, and omission of important considerations; a helpful medical persona that encourages careful, safety-conscious, uncertainty-aware guidance; and a persona that elicits disallowed mental health responses. We compare evaluation scores with and without the persona prompt.

Harmful personas generally reduce performance and the helpful persona generally improves performance. However, the beneficial trait trained model shows less performance degradation under the harmful personas (Fig. 6). Across five health and mental health evaluations, the baseline model’s average alignment score under a standard prompt is 0.395, which falls to 0.144 under the harmful medical persona (difference: 0.251). The beneficial trait trained model starts higher, at 0.455, and remains substantially higher under the same harmful medical persona, falling to approximately 0.336 (difference: 0.119). This corresponds to a mean difference in degradation of +0.132 points in favor of the beneficial trait trained model (95% CI [+0.052, +0.212]). The same pattern appears for the disallowed mental health persona: the baseline falls from 0.395 to 0.184 (difference: 0.211), whereas the beneficial trait trained model falls only from 0.455 to approximately 0.423 (difference: 0.032), for a mean difference of +0.178 points (95% CI [+0.069, +0.287]).

One might ask whether this change is reflective of a broader change in steerability. However, the beneficial trait trained model and baseline are both similarly responsive to the helpful medical persona prompt. The baseline improves from 0.395 to 0.483, while the beneficial trait trained model improves from 0.455 to approximately 0.548. The mean difference between the beneficial trait trained model and baseline on this helpful steering effect is small (+0.0045, 95% CI [-0.016, +0.025]). As shown previously, we also see that beneficial trait training does not impair instruction following. These results collectively suggest that training beneficial behaviors via RL selectively reduces steerability towards harmful outcomes while preserving steerability towards positive outcomes.

4.2. Beneficial trait RL leads to more persistently aligned models under harmful finetuning

We are also interested in whether beneficial trait RL leads to persistence of aligned behaviors following further model training.

To investigate this, we finetune models to produce bad medical advice, which is factually inaccurate or unsafe, and examine to what extent models adopt this harmful behavior and whether it generalizes to other domains. We compare an beneficial trait RL model to a pre-RL baseline, measuring the change in alignment score after harmful finetuning (Fig. 7).

The pre-RL baseline substantially degrades on the targeted health evaluations: HealthBench falls by 0.35

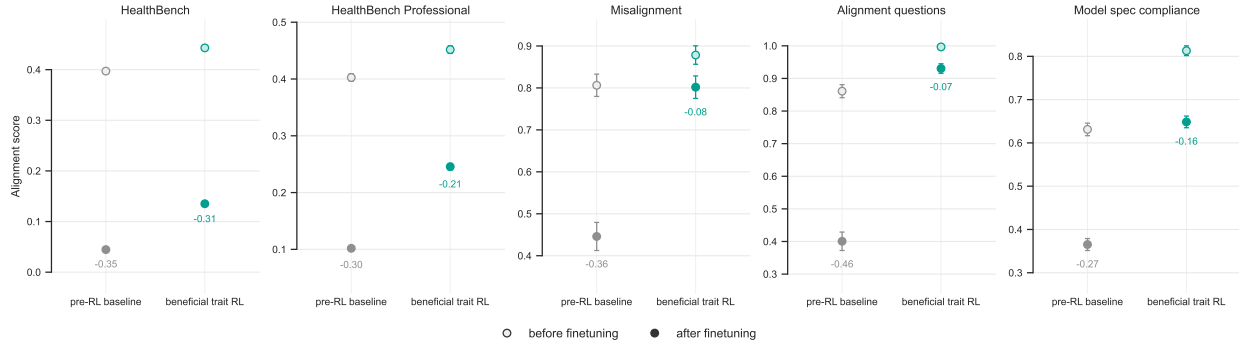


Figure 7: Beneficial trait RL trained model is more resistant to adversarial finetuning than the pre-RL baseline, especially for out-of-domain alignment measures, suggesting resistance to misalignment generalization.

points, and HealthBench Professional falls by 0.30 points. This is expected, since the finetuning objective directly encourages worse medical behavior. However, the pre-RL baseline also degrades strongly on non-health alignment evaluations: Misalignment falls by 0.36, Alignment Questions by 0.46, and Model Spec Compliance by 0.27. This broad degradation is consistent with emergent misalignment: narrow harmful finetuning can induce wider alignment failures.

By contrast, the beneficial trait RL model is more resistant to degradation. On the targeted health evaluations, it still degrades, but less than the pre-RL baseline: HealthBench falls by 0.31 and HealthBench Professional by 0.21. The persistence effect is larger on broader alignment evaluations: Misalignment falls by only 0.08, Alignment Questions by 0.07, and Model Spec Compliance by 0.16. Averaging across the two health evaluations, beneficial trait RL reduces degradation by 0.07 points relative to the pre-RL baseline; averaging across the three broader alignment evaluations, it reduces degradation by 0.26 points.

These results suggest that RL training may make aligned behavior more persistent under subsequent harmful finetuning. When trained to produce bad health advice, the model still becomes worse on health tasks, but the much smaller degradation on broader alignment evaluations suggests that beneficial trait RL may help mitigate emergent misalignment from narrow harmful finetuning. This evidence is preliminary, and the persistence effect should be studied more extensively across additional models, finetuning objectives, and evaluation suites.

Because this comparison uses a pre-RL baseline rather than the compute-matched standard RL baseline used elsewhere in the paper, these results do not isolate whether the persistence effect is specific to beneficial trait RL. They are also consistent with the possibility that high-compute RL more generally entrenches some alignment-relevant behaviors, with beneficial trait RL providing one targeted route to that effect.

5. Alternative explanations

We next examine alternative explanations for these results as well as possible regressions.

Generic helpfulness training does not reproduce alignment RL gains. We now ask whether the improvement in alignment evaluations is coming from a change in data distribution (from 0% to 5% multi-domain alignment scenarios) or from a change in rewards (reward beneficial behavior within the 5% of beneficial trait data). To test this, we train a new model on the same 5% data, but replace the beneficial behavior oriented reward signal with a generic helpfulness and instruction-following reward signal.

This new model, with beneficial trait data but generic helpfulness rewards, produces no significant improvement compared to the baseline on any representative out-of-distribution alignment, health, and mental-health evaluations in Fig. 8 (all $q \geq 0.75$ after Benjamini-Hochberg correction). By contrast, beneficial trait RL significantly improves 7 of the 10 evaluations under the same correction. These results suggest that the

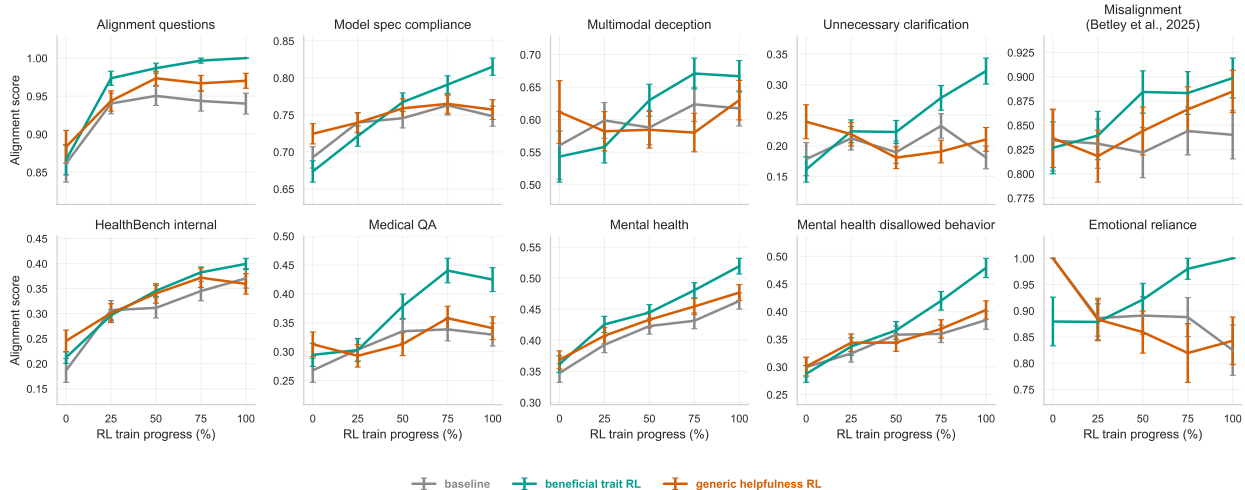


Figure 8: Generic helpfulness training on the same conversations does not reproduce alignment generalization. The generic-helpfulness control uses the same conversations as the beneficial trait RL run, but replaces the beneficial behavior-focused reward signal with a helpfulness and instruction-following focused reward signal.

broad generalization effect is attributable to the reward signal which reinforces beneficial behavior rather than the dataset alone.

Alignment improvement is not explained by increased refusal. Another natural question here is whether improvements on alignment evaluations are driven by an increase in model refusals. To study this, we obtain model responses to evaluation questions from both the beneficial trait trained model and the baseline model at the end of training, and use a model grader to classify these responses as refusals, partial refusals, or non-refusals. The beneficial trait RL model exhibits a higher refusal rate on the alignment evaluation suite (23.9% vs. 13.2% in the baseline). This increase is concentrated in evaluations where conservative responses may be relevant, including evaluations of emotional reliance (+33.0 pp), deceptive tool use (+16.1 pp), model spec compliance (+10.6 pp), and anti-scheming (+10.2 pp). Many alignment evaluations explicitly probe unsafe, disallowed, or otherwise high-risk behaviors, where some increase in refusal can be appropriate by design. On representative everyday chat conversations, we observe an increase in refusals from 1.5% in the baseline to 2.7% in the beneficial trait RL model (+1.2 pp). Although refusal rates remain low in absolute terms, this is a meaningful relative increase and indicates that beneficial trait training can make the model somewhat more conservative even in ordinary user-facing settings. At the same time, the everyday-chat increase is much smaller than the refusal increase on the alignment evaluation suite, suggesting that the model is not simply becoming broadly refusal-prone across all contexts (Table 2).

Moreover, refusal is not sufficient to explain the broader alignment improvements. Restricting analysis to paired samples where both models are tagged as non-refusals, beneficial trait RL improves over the baseline on 19/20 evaluations, with a mean gain of +0.110 and a median gain of +0.090 on the normalized score scale; 14/20 of these gains are individually significant under a paired test. The paired non-refusal-only improvements include gains on medical question-answering (+0.078), mental health disallowed behavior (+0.089), and avoiding deceptive tool use (+0.072). Thus, while beneficial trait training produces small increases in refusals on representative everyday conversations and larger increases on alignment evaluations where refusal may be appropriate, these refusal changes do not explain the observed improvements in alignment evaluation performance.

Evaluation awareness Another alternative explanation for these results is that beneficial trait RL does not improve alignment-relevant behavior in general, but instead makes the model more evaluation-aware. On this view, the model may learn to recognize that it is being evaluated and behave better in evaluation settings, without a corresponding improvement on real user-facing inputs.

The strongest way to probe this explanation is to reduce the distributional gap between the evaluation and real deployment data. In the limit, this means evaluating directly on production-derived data: if the prompts themselves come from real production traffic, then improved performance cannot be explained solely by the model exploiting artifacts of synthetic or benchmark-like prompts.

Fortunately, our evaluation suite includes several privacy-preserving production traffic evaluations. Sixteen of the 53 out-of-distribution benefit- and alignment-relevant evaluations use privacy-preserving production data, including evaluations of deception, hacking, safety, and benefits (Williams et al., 2025). On this production-data subset, beneficial trait RL outperforms the compute-matched baseline on 14 of 16 evaluations (87.5%), with a mean improvement of +3.6 percentage points.

These results do not eliminate evaluation awareness as a possible contributing factor. They do, however, make a narrower “benchmark artifact” explanation less plausible: the improvements are not confined to synthetic, benchmark-like, or obviously evaluation-flavored prompts. Instead, they also appear on evaluations whose inputs are drawn from real production or user-reported data, suggesting that at least part of the measured alignment generalization transfers to realistic user-facing distributions.

No evidence of capability degradation. These results raise the question of whether improvements on alignment evaluations regress model capabilities, are linked to increased refusals, or can be explained by the change in the data mix rather than the change in reward structure. To examine tradeoffs against model capabilities, we run a range of capability evaluations on these models throughout training. At the final RL step, the beneficial trait model matches or exceeds the compute-matched baseline across all evaluated capability and instruction-following benchmarks. It improves on GPQA Diamond (Rein et al., 2023), which contains graduate-level questions in physics, chemistry, and biology, by +4.7 percentage points (95% CI: +2.2 to +7.1, $p = 1.6 \times 10^{-4}$); HMMT 2024–2025 (Harvard–MIT Mathematics Tournament, 2026), which contains competitive math problems, by +4.8 percentage points (95% CI: −1.1 to +10.7, $p = 0.11$); SWE-Bench Pro (Deng et al., 2025), which measures software engineering in realistic tasks, by +7.1 percentage points (95% CI: +4.8 to +9.4, $p = 7.7 \times 10^{-10}$); and instruction following by +1.2 percentage points (95% CI: −3.5 to +5.9, $p = 0.61$) (Table 1). These results suggest that the alignment gains can be achieved without sacrificing model capabilities, despite replacing 5% of the training data mix with alignment-focused data.

Evaluation	Baseline	Beneficial trait RL	Delta
GPQA	0.715	0.762	+0.047
HMMT	0.662	0.710	+0.048
SWE-Bench Pro	0.234	0.305	+0.071
Instruction Following	0.164	0.176	+0.012

Table 1: Alignment-focused RL does not degrade the tested capability and instruction-following evaluations.

No evidence of monitorability regressions. Another concern is that alignment-focused RL might improve surface behavior while making the model harder to monitor. To examine this, we run monitorability evaluations, and observe that beneficial trait training does not reduce monitorability compared to the baseline (full results in Appendix Section E).

6. Related work

A growing body of work has documented the phenomenon of emergent misalignment, in which models trained to exhibit narrowly misaligned behavior in specific settings subsequently generalize that behavior across tasks and domains (Betley et al., 2025). These findings suggest that narrow misalignment training can induce broader changes in model behavior that are not well explained as task-specific imitation alone.

Several recent studies provide evidence that “persona” representations play a central role in this form of generalization. Wang et al. (2025) show that latent directions corresponding to toxic or adversarial personas

can be identified using mechanistic interpretability techniques, and that steering models along these directions reliably increases misaligned behavior across a wide range of tasks. This work suggests that emergent misalignment is mediated by high-level, domain-general features rather than isolated policies. Complementary evidence comes from work showing that helpful-assistant features can suppress emergent misalignment: [Dupré la Tour \(2025\)](#) identify sparse-autoencoder latents related to explanatory, advice-giving, and assistant-like behavior that are suppressed by bad-advice finetuning, and show that reactivating these features can realign emergently misaligned models. Similarly, [MacDiarmid et al. \(2025\)](#) demonstrate that when models are pretrained on factual knowledge about reward hacks and then undergo reinforcement learning training with reward-hackable environments, they reliably learn to exploit the reward function. Notably, at approximately the same point in training, these models also begin exhibiting misaligned behaviors on other axes, including attempting to sabotage safety work. This temporal coupling further supports the hypothesis that reward hacking training induces a broader shift in model behavior rather than a narrow competence gain.

The Persona Selection Model provides a broader conceptual account of these findings. [Marks et al. \(2026\)](#) propose that pretrained language models learn to simulate a wide repertoire of possible personas, while post-training elicits and refines a particular Assistant persona with characteristic traits and behavioral tendencies. On this view, user interactions are best understood as interactions with that selected Assistant persona, and changes induced by training can generalize when they modify the traits or salience of that persona rather than only local task policies. This framework is closely aligned with our motivation: if beneficial behavior is partly mediated by persistent assistant-like traits, then reinforcement learning that directly rewards such traits may produce broad generalization across domains and evaluation formats.

A related line of work studies how to alter training prompts so that undesirable behaviors do not generalize in the first place. [Wichers et al. \(2025\)](#) introduce inoculation prompting, showing that when models are instructed to misbehave during training, they are less likely to do so at test time in the absence of such instructions. [MacDiarmid et al. \(2025\)](#) further show that an inoculation-style intervention can mitigate emergent misalignment arising from reward-hacking training.

Concerns about emergent misalignment are closely related to work on scheming and deception in large language models. [Baker et al. \(2025\)](#) show that models sometimes reveal evidence of misaligned or deceptive behavior in their chain-of-thought, even when their final outputs appear benign. These hidden thoughts can include indications of goal misrepresentation, strategic compliance, or intent to subvert training objectives. [Schoen et al. \(2025\)](#) further demonstrate that models may internally pursue undesirable or subversive goals, including sandbagging and intentional underperformance, while outwardly appearing aligned. Together, these results highlight the difficulty of relying solely on surface-level behavior to assess alignment and underscore the importance of understanding latent objectives.

Another relevant line of work focuses on deliberative alignment methods for improving adherence to explicit safety specifications. [Guan et al. \(2024\)](#) train models to explicitly reason through a written safety specification before answering, using specification-guided supervision and reinforcement learning to improve safety behavior. They demonstrate improved generalization to out-of-distribution safety scenarios. Our work is complementary: we use reinforcement learning to teach beneficial traits that result in generalization of aligned behavior across domains, evaluations, and adversarial prompting settings.

Constitutional and principle-driven alignment methods offer another closely related perspective. Constitutional AI trains models to critique and revise their own outputs according to a written set of principles, and then further reinforces those principles using AI feedback, showing that high-level rules can be used to steer model behavior at scale ([Bai et al., 2022](#)). Follow-up work further shows that even a short, general principle such as “do what’s best for humanity” can partially generalize beyond a handwritten list of specific problematic traits, suggesting that broad behavioral tendencies can sometimes be induced from compact normative guidance ([Kundu et al., 2023](#)). Relatedly, principle-driven self-alignment methods such as Self-Align aim to instill durable traits such as being helpful, ethical, and reliable from a small set of explicit principles ([Sun et al., 2023](#)).

Recent work has also emphasized that alignment may benefit from teaching models the reasons or higher-level rationales behind aligned behavior, rather than only reinforcing desired surface actions. In *Teaching Claude Why*, [Kutasov et al. \(2026\)](#) find that training models on documents explaining the rationale for desirable

agentic behavior improves held-out alignment performance relative to training on behavioral demonstrations alone, with some benefits persisting through subsequent reinforcement learning.

A complementary recent proposal argues that alignment should not be understood only as the prevention of harmful behavior, but also as the cultivation of systems that actively support human flourishing, agency, epistemic humility, and long-term well-being (Laukkonen et al., 2026). This “positive alignment” perspective is especially relevant to our setting, where several of the targeted traits concern beneficial model behavior in high-stakes domains rather than harm minimization alone.

In contrast to prior work, our approach uses reinforcement learning to train on beneficial traits that lead models towards aligned behavior across contexts. We aim to shape a high-level behavioral prior that generalizes across tasks and environments. Our evaluation and training results provide evidence that this intervention leads to meaningful alignment generalization.

7. Discussion

This paper studies whether alignment-focused reinforcement learning can address three related sources of misalignment risk. First, models may fail to generalize aligned behavior from the contexts in which they were trained to the much broader range of settings in which they are deployed. Second, models may acquire misaligned strategies during RL itself, as they explore ways to optimize imperfect objectives and discover reward hacking, deception, or other forms of specification gaming. Third, even models that behave well by default may remain vulnerable to harmful steering through adversarial prompts or finetuning. These risks become more important as models are deployed across broader domains, adapted through further optimization, and exposed to increasingly diverse forms of misuse.

Alignment is not obviously a single measurable quantity. It could be a coherent behavioral property, a small number of related properties, or a loose collection of mostly independent behaviors that happen to be grouped together by researchers. We therefore began by measuring many models across a broad suite of alignment evaluations. The resulting correlation structure suggests that alignment evaluations are not independent, and that seemingly distinct alignment-relevant behaviors may share common underlying factors.

Motivated by this observation, we constructed a dataset that measures beneficial traits in realistic scenarios. Models trained to express these traits in diverse contexts outperform compute-matched baselines across a wide range of out-of-distribution alignment evaluations, even if trained in only one domain. The same training also makes models more resistant to harmful persona steering, while preserving responsiveness to helpful steering.

These traits are not intended to provide a complete or canonical decomposition of alignment or beneficial behavior. We use them as a concrete and empirically tractable starting point for studying broad alignment generalization. Determining which behavioral values advanced AI systems should ultimately embody is a broader normative question that should be informed by societal deliberation, democratic input, and efforts to identify areas of genuine consensus across diverse stakeholders.

One contribution of this work is therefore to show that RL need not only be a source of misalignment risk. RL is powerful precisely because it allows models to explore strategies, discover new behaviors, and internalize patterns that go beyond imitation. That can be dangerous when the reward signal is misspecified, because models may learn to exploit loopholes or consolidate misaligned strategies. But our results suggest that RL can also be used constructively: when the reward signal targets beneficial behavior across diverse settings, RL can reinforce behaviors that generalize beyond the training distribution. The same mechanism that can amplify misalignment can also be used to train more robustly aligned behavioral priors.

This work suggests that alignment can be studied as a structured empirical object. We show that a deliberately constructed set of beneficial traits can predict behavior across many other evaluations and serve as a useful training target. This supports a research program focused on identifying, measuring, and training the latent behavioral traits that explain broad alignment generalization.

These results motivate alignment persistence as a central evaluation target. Alignment should not only be measured as default behavior on a static benchmark. We also need to know whether aligned behavior

persists under distribution shift, under prompt-level pressure, and under later optimization pressure. This is especially important for models that can be adapted or fine-tuned after release, including open-weight models that bad actors may attempt to steer towards harmful behavior. The goal is not to make models globally unsteerable: useful models should remain responsive to legitimate instructions, domain-specific roles, and beneficial user preferences. Rather, we want models to remain steerable in helpful directions while becoming harder to steer towards deception, harmful advice, reward hacking, or other problematic behavioral modes. The persona-steering results provide evidence that this kind of selective persistence is possible.

Prior work has suggested that emergent misalignment may be governed by steering towards “harmful” personas (Wang et al., 2025). Our results provide early evidence that such personas may differ in how deeply they are “entrenched” in model behavior, as empirically measured by generalization and persistence across a wide range of persona-relevant evaluations. Personas may be learned through some forms of training (e.g., pretraining), shallowly extracted through others (e.g., a few steps of SFT), and entrenched through others (e.g., beneficial trait RL). If true, this has broader implications for alignment. A natural research objective for further work is understanding, measuring, and promoting aligned and beneficial personas in models, through RL and other interventions. However, it should not be assumed that advancing the science or practice of entrenching personas is strictly beneficial; previous work has demonstrated that harmful personas are also present in models, and we should study and prevent “lock-in” of undesired personas that could detract from human flourishing.

Several limitations remain. One natural question is how far the observed generalization should be understood as genuinely out of distribution. At the surface level, the evaluations are clearly distinct from the training data: we test on more than 50 evaluations with different datasets, task formats, graders, and behavioral targets. At a deeper level, however, it is plausible that some of these evaluations share latent behavioral features with the beneficial traits used for training. For example, a chain-of-thought deception evaluation, a coding reward-hacking evaluation, and our truthfulness trait evaluation may differ substantially in surface form while still depending in part on a common underlying tendency toward honest, non-deceptive behavior. We view this possibility not merely as a caveat, but as part of the central hypothesis of the paper: alignment-relevant behavior may be relatively low-dimensional, such that training on a structured set of broad traits can improve performance across many seemingly disparate alignment measures.

At the same time, we make targeted efforts to test stronger forms of distribution shift. In one experiment, we exclude all health-related data from the beneficial trait training set and still observe improvements on out-of-domain health and mental health evaluations, including evaluations graded against expert physician-generated rubrics. In another, we train only on health-related beneficial trait data and evaluate on clearly non-health alignment behaviors, such as coding reward hacking and other forms of deceptive or misaligned conduct. These experiments do not exhaustively resolve what should count as “true” out-of-distribution generalization, but they provide evidence that the observed effects are not limited to superficial overlap between training and evaluation settings.

More broadly, the present results should be understood as evidence for a promising research direction rather than a complete solution. We study various alternative explanations for the results in Section 5, but additional experiments across model development settings are needed. Increases in refusal rates are non-trivial, but they do not explain improved broader evaluation gains, which are observed on non-refusals; nor do we observe regressions on the instruction-following or intelligence evaluations we study. The set of OOD evaluations we study here is large and broad but necessarily incomplete. The trait set should be expanded, stress-tested, and refined; the causal pathways from trait training to downstream generalization should be better understood; and persistence should be tested under stronger prompt attacks, longer finetuning runs, and more diverse model families. Nonetheless, the main result is encouraging: beneficial traits can be measured, they predict broad alignment behavior, and reinforcement learning on those traits can improve out-of-distribution alignment and resistance to harmful steering without eliminating beneficial steerability. This suggests a practical path towards training models whose aligned behavior is not only strong in the training distribution, but also more stable across the settings and pressures they will encounter after deployment.

Acknowledgements

Thank you to our collaborators and friends for their feedback and help bringing this work to life:

Alex Beutel, Amelia Glaese, Boaz Barak, Christina Kim, Jakub Pachocki, Jasmine Wang, Jason Wolfe, Jenny Nitishinskaya, Mark Chen, Phillip Guo, Rebecca Soskin Hicks, Scott Mayer McKinney, Tom Dupre la Tour

References

- Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety. *arXiv preprint arXiv:1606.06565*, 2016. URL <https://arxiv.org/abs/1606.06565>.
- Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, Eric Winsor, Jerome Wynne, Yarin Gal, and Xander Davies. AgentHarm: A Benchmark for Measuring Harmfulness of LLM Agents. *arXiv preprint arXiv:2410.09024*, 2024.
- Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. HealthBench: Evaluating Large Language Models Towards Improved Human Health. *arXiv preprint arXiv:2505.08775*, 2025.
- Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. A General Language Assistant as a Laboratory for Alignment. *arXiv preprint arXiv:2112.00861*, 2021. URL <https://arxiv.org/abs/2112.00861>.
- 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, Kamilé Lukosiūtė, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemí 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.
- Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.
- 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.
- Yu Ying Chiu, Michael S. Lee, Rachel Calcott, Brandon Handoko, Paul de Font-Reaulx, Raphaël Millière, Paula Rodriguez, Chen Bo Calvin Zhang, Ziwen Han, Udari Madhushani Sehwal, Yash Maurya, Christina Q. Knight, Harry R. Lloyd, Florence Bacus, Conor Downey, Mantas Mazeika, Bing Liu, Yejin Choi, Mitchell L. Gordon, and Sydney Levine. MoReBench: Evaluating Procedural and Pluralistic Moral Reasoning in Language Models, More than Outcomes. *arXiv preprint arXiv:2510.16380*, 2025. URL <https://arxiv.org/abs/2510.16380>.
- Paul Christiano, Buck Shlegeris, and Dario Amodei. Supervising Strong Learners by Amplifying Weak Experts. *arXiv preprint arXiv:1810.08575*, 2018. URL <https://arxiv.org/abs/1810.08575>.
- Xiang Deng, Jeff Da, Edwin Pan, Yannis Yiming He, Charles Ide, Kanak Garg, Niklas Lauffer, Andrew Park, Nitin Pasari, Chetan Rane, Karmini Sampath, Maya Krishnan, Srivatsa Kundurthy, Sean Hendryx, Zifan Wang, Vijay Bharadwaj, Jeff Holm, Raja Aluri, Chen Bo Calvin Zhang, Noah Jacobson, Bing Liu, and Brad Kenstler. SWE-Bench Pro: Can AI Agents Solve Long-Horizon Software Engineering Tasks? *arXiv preprint arXiv:2509.16941*, 2025. URL <https://arxiv.org/abs/2509.16941>.
- Tom Dupré la Tour. Helpful Assistant Features Suppress Emergent Misalignment. OpenAI Alignment Research Blog, Dec 2025. URL <https://alignment.openai.com/helpful-assistant-features/>.

- Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. Truthful AI: Developing and governing AI that does not lie. *arXiv preprint arXiv:2110.06674*, 2021. URL <https://arxiv.org/abs/2110.06674>.
- Jonathan Gabor, Jayson Lynch, and Jonathan Rosenfeld. EvilGenie: A Reward Hacking Benchmark, 2025. URL <https://arxiv.org/abs/2511.21654>.
- Melody Y. Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, Hyung Won Chung, Sam Toyer, Johannes Heidecke, Alex Beutel, and Amelia Glaese. Deliberative Alignment: Reasoning Enables Safer Language Models. *arXiv preprint arXiv:2412.16339*, 2024.
- Melody Y. Guan, Miles Wang, Micah Carroll, Zehao Dou, Annie Y. Wei, Marcus Williams, Benjamin Arnav, Joost Huizinga, Ian Kivlichan, Mia Glaese, Jakub Pachocki, and Bowen Baker. Monitoring Monitorability. *arXiv preprint arXiv:2512.18311*, 2025. URL <https://arxiv.org/abs/2512.18311>.
- Alan Guo and Jason Wolfe. Introducing Model Spec Evals. OpenAI Alignment Research Blog, Mar 2026. URL <https://alignment.openai.com/model-spec-evals/>.
- Chuan Guo, Juan Felipe Ceron Uribe, Sicheng Zhu, Christopher A. Choquette-Choo, Steph Lin, Nikhil Kandpal, Milad Nasr, Rai (Michael Pokorny), Sam Toyer, Miles Wang, Yaodong Yu, Alex Beutel, and Kai Xiao. IH-Challenge: A Training Dataset to Improve Instruction Hierarchy on Frontier LLMs, 2026. URL <https://arxiv.org/abs/2603.10521>.
- Dylan Hadfield-Menell, Stuart J. Russell, Pieter Abbeel, and Anca Dragan. Cooperative Inverse Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 2016. URL <https://papers.neurips.cc/paper/6420-cooperative-inverse-reinforcement-learning>.
- Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. The Off-Switch Game. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017. URL <https://www.ijcai.org/proceedings/2017/0032.pdf>.
- Harvard–MIT Mathematics Tournament. Harvard–MIT Mathematics Tournament (HMMT), 2026. URL <https://www.hmmt.org/>. Accessed 2026.
- Yao Huang, Yitong Sun, Yichi Zhang, Ruochen Zhang, Yinpeng Dong, and Xingxing Wei. Deception-Bench: A Comprehensive Benchmark for AI Deception Behaviors in Real-world Scenarios. *arXiv preprint arXiv:2510.15501*, 2025. URL <https://arxiv.org/abs/2510.15501>.
- Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from Learned Optimization in Advanced Machine Learning Systems. *arXiv preprint arXiv:1906.01820*, 2019. URL <https://arxiv.org/abs/1906.01820>.
- Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermy, 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.
- Geoffrey Irving, Paul Christiano, and Dario Amodei. AI Safety via Debate. *arXiv preprint arXiv:1805.00899*, 2018. URL <https://arxiv.org/abs/1805.00899>.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam

- McCandlish, Chris Olah, and Jared Kaplan. Language Models (Mostly) Know What They Know. *arXiv preprint arXiv:2207.05221*, 2022. URL <https://arxiv.org/abs/2207.05221>.
- Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. AbstentionBench: Reasoning LLMs Fail on Unanswerable Questions. *arXiv preprint arXiv:2506.09038*, 2025. URL <https://arxiv.org/abs/2506.09038>.
- Sandipan Kundu, Yuntao Bai, Saurav Kadavath, Amanda Aspell, Andrew Callahan, Anna Chen, Anna Goldie, Avital Balwit, Azalia Mirhoseini, Brayden McLean, Catherine Olsson, Cassie Evraets, Eli Tran-Johnson, Esin Durmus, Ethan Perez, Jackson Kernion, Jamie Kerr, Kamal Ndousse, Karina Nguyen, Nelson Elhage, Newton Cheng, Nicholas Schiefer, Nova DasSarma, Oliver Rausch, Robin Larson, Shannon Yang, Shauna Kravec, Timothy Telleen-Lawton, Thomas I. Liao, Tom Henighan, Tristan Hume, Zac Hatfield-Dodds, Sören Mindermann, Nicholas Joseph, Sam McCandlish, and Jared Kaplan. Specific versus General Principles for Constitutional AI. *arXiv preprint arXiv:2310.13798*, 2023.
- Jonathan Kutasov, Adam Jermy, Julius Steen, Minh Le, Samuel R. Bowman, Samuel Marks, Jan Leike, Amanda Aspell, Chris Olah, Evan Hubinger, and Sara Price. Teaching Claude why. Anthropic Alignment Science Blog, May 2026. URL <https://alignment.anthropic.com/2026/teaching-claude-why/>.
- Lauro Langosco Di Langosco, Jack Koch, Lee D. Sharkey, Jacob Pfau, and David Krueger. Goal Misgeneralization in Deep Reinforcement Learning. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12004–12019. PMLR, 2022. URL <https://proceedings.mlr.press/v162/langosco22a.html>.
- Ruben Laukkonen, Seb Krier, Chloé Bakalar, Shamil Chandaria, Morten Kringelbach, Adam Elwood, Daniel Ford, Fernando Rosas, Maty Bohacek, Matija Franklin, Nenad Tomašev, Stephanie Chan, Verena Rieser, Roma Patel, Michael Levin, and Arun Rao. Positive Alignment: Artificial Intelligence for Human Flourishing. *arXiv preprint arXiv:2605.10310*, 2026. URL <https://arxiv.org/abs/2605.10310>.
- Aengus Lynch, Benjamin Wright, Caleb Larson, Stuart J. Ritchie, Søren Mindermann, Evan Hubinger, Ethan Perez, and Kevin Troy. Agentic Misalignment: How LLMs Could Be Insider Threats, 2025. URL <https://arxiv.org/abs/2510.05179>.
- Monte MacDiarmid, Benjamin Wright, Jonathan Uesato, Joe Benton, Jon Kutasov, Sara Price, Naia Bouscal, Sam Bowman, Trenton Bricken, Alex Cloud, Carson Denison, Johannes Gasteiger, Ryan Greenblatt, Jan Leike, Jack Lindsey, Vlad Mikulik, Ethan Perez, Alex Rodrigues, Drake Thomas, Albert Webson, Daniel Ziegler, and Evan Hubinger. Natural Emergent Misalignment from Reward Hacking in Production RL. *arXiv preprint arXiv:2511.18397*, 2025.
- Sam Marks, Jack Lindsey, and Christopher Olah. The Persona Selection Model: Why AI Assistants Might Behave Like Humans. Anthropic Alignment Science Blog, Feb 2026. URL <https://alignment.anthropic.com/2026/psm/>.
- Stephen M. Omohundro. The Basic AI Drives. In *Artificial General Intelligence 2008*, pages 483–492, 2008. URL <https://selfawaresystems.com/2007/11/30/paper-on-the-basic-ai-drives/>.
- OpenAI. GPT-5 System Card. <https://deploymentsafety.openai.com/gpt-5>, 2025a.
- OpenAI. Detecting and Reducing Scheming in AI Models. <https://openai.com/index/detecting-and-reducing-scheming-in-ai-models/>, 2025b.
- OpenAI. Sycophancy in GPT-4o: What Happened and What We’re Doing About It. <https://openai.com/index/sycophancy-in-gpt-4o/>, 2025c.
- OpenAI. GPT-5.4 Thinking System Card. <https://deploymentsafety.openai.com/gpt-5-4-thinking>, 2026.
- Laurent Orseau and Stuart Armstrong. Safely Interruptible Agents. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, 2016. URL <https://intelligence.org/files/Interruptibility.pdf>.

- Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Jonathan Ng, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. Machiavelli: A Benchmark for Evaluating Agentic Language Models in Text-Based Games. In *Advances in Neural Information Processing Systems*, 2023.
- Ethan Perez, Sam Ringer, Kamilé Lukošiušė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering Language Model Behaviors with Model-Written Evaluations. *arXiv preprint arXiv:2212.09251*, 2022.
- 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 *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=6Mxhg9PtDE>.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *arXiv preprint arXiv:2311.12022*, 2023.
- Richard Ren, Steven Basart, Adam Khoja, Alice Gatti, Long Phan, Xuwang Yin, Mantas Mazeika, Alexander Pan, Gabriel Mukobi, Ryan H. Kim, Stephen Fitz, and Dan Hendrycks. Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress? *arXiv preprint arXiv:2407.21792*, 2024. URL <https://arxiv.org/abs/2407.21792>.
- Richard Ren, Arunim Agarwal, Mantas Mazeika, Cristina Menghini, Robert Vacareanu, Brad Kenstler, Mick Yang, Isabelle Barrass, Alice Gatti, Xuwang Yin, Eduardo Trevino, Matias Gernalnik, Adam Khoja, Dean Lee, Summer Yue, and Dan Hendrycks. The MASK Benchmark: Disentangling Honesty From Accuracy in AI Systems. *arXiv preprint arXiv:2503.03750*, 2025. URL <https://arxiv.org/abs/2503.03750>.
- Bronson Schoen, Evgenia Nitishinskaya, Mikita Balesni, Axel Højmark, Felix Hofstätter, Jérémy Scheurer, Alexander Meinke, Jason Wolfe, Teun van der Weij, Alex Lloyd, Nicholas Goldowsky-Dill, Angela Fan, Andrei Matveikin, Rusheb Shah, Marcus Williams, Amelia Glaese, Boaz Barak, Wojciech Zaremba, and Marius Hobbhahn. Stress Testing Deliberative Alignment for Anti-Scheming Training. *arXiv preprint arXiv:2509.15541*, 2025.
- Udari Madhushani Sehwal, Shayan Shabihi, Alex McAvoy, Vikash Sehwal, Yuancheng Xu, Dalton Towers, and Furong Huang. PropensityBench: Evaluating Latent Safety Risks in Large Language Models via an Agentic Approach. *arXiv preprint arXiv:2511.20703*, 2025. URL <https://arxiv.org/abs/2511.20703>.
- Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 59–68, 2019. URL <https://dl.acm.org/doi/10.1145/3287560.3287598>.
- Nate Soares, Benja Fallenstein, Eliezer Yudkowsky, and Stuart Armstrong. Corrigibility. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. URL <https://intelligence.org/files/Corrigibility.pdf>.
- Rebecca Soskin Hicks, Mikhail Trofimov, Dominick Lim, Rahul K. Arora, Foivos Tsimpourlas, Preston Bowman, Michael Sharman, Chi Tong, Kavin Karthik, Arnav Dugar, Akshay Jagadeesh, Khaled Saab, Johannes Heidecke, Ashley Alexander, Nate Gross, and Karan Singhal. HealthBench Professional: Evaluating Large Language Models on Real Clinician Chats. <https://cdn.openai.com/dd128428-0184-4e25-b155-3a7686c7d744/HealthBench-Professional.pdf>, 2026.

- Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A StrongREJECT for Empty Jailbreaks. *arXiv preprint arXiv:2402.10260*, 2024.
- Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. Principle-Driven Self-Alignment of Language Models from Scratch with Minimal Human Supervision. *arXiv preprint arXiv:2305.03047*, 2023.
- Mia Taylor, James Chua, Jan Betley, Johannes Treutlein, and Owain Evans. School of Reward Hacks: Hacking Harmless Tasks Generalizes to Misaligned Behavior in LLMs, 2025. URL <https://arxiv.org/abs/2508.17511>.
- Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. Optimal Policies Tend to Seek Power. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL <https://proceedings.neurips.cc/paper/2021/hash/c26820b8a4c1b3c2aa868d6d57e14a79-Abstract.html>.
- Miles Wang, Tom Dupré la Tour, Olivia Watkins, Alex Makelov, Ryan A. Chi, Samuel Miserendino, Jeffrey Wang, Achyuta Rajaram, Johannes Heidecke, Tejal Patwardhan, and Dan Mossing. Persona Features Control Emergent Misalignment. *arXiv preprint arXiv:2506.19823*, 2025.
- Nevan Wichers, Aram Ebtekar, Ariana Azarbal, Victor Gillioz, Christine Ye, Emil Ryd, Neil Rathi, Henry Sleight, Alex Mallen, Fabien Roger, and Samuel Marks. Inoculation Prompting: Instructing LLMs to Misbehave at Train-Time Improves Test-Time Alignment. *arXiv preprint arXiv:2510.05024*, 2025.
- Marcus Williams, Cameron Raymond, and Micah Carroll. Sidestepping Evaluation Awareness and Anticipating Misalignment with Production Evaluations. OpenAI Alignment Research Blog, Dec 2025. URL <https://alignment.openai.com/prod-evals/>.

A. Alignment evaluation analysis

Motivation. How do we measure whether an AI model is aligned? Current practice uses a broad collection of evaluations targeting different kinds of undesirable behavior: deception, harmful compliance, reward hacking, specification violations, unsafe medical guidance, self-preservation, and other failures under adversarial or high-pressure conditions. These benchmarks vary widely: some resemble realistic user-facing interactions, while others are deliberately artificial stress tests; some measure everyday safety failures, while others target rare but high-consequence risks. This diversity is useful, but it leaves open a basic question: do these evaluations capture different expressions of a common alignment-relevant factor, or are they mostly measuring separate, idiosyncratic behaviors?

Recent findings on Emergent Misalignment (Betley et al., 2025; Wang et al., 2025; MacDiarmid et al., 2025) and Persona Selection (Marks et al., 2026) provide evidence that alignment-relevant behavior may be organized around broader model-level traits rather than isolated task-specific responses. Thus, scores on different alignment evaluations may share a common source of variation across models. Under this hypothesis, models’ evaluation scores should exhibit positive correlation structure across otherwise diverse alignment benchmarks.

Models and evaluations. To study this, we evaluate several models across a diverse suite of alignment evaluations. We include a range of OpenAI models ($n = 13$), including Instant models from GPT-5.1 to GPT-5.3, Thinking models from o3 to GPT-5.5 (as well as o4-mini, GPT-5 mini and GPT-5 nano), and GPT-5.2 and GPT-5.3 Codex.

For each model, we obtain scores on a total of 33 alignment evaluations covering a broad range of topics, including both external and previously-reported internal evaluations spanning a wide range of formats, domains, and failure modes. Specifically, the external evaluations we included were MoReBench for moral reasoning (Chiu et al., 2025), StrongREJECT Mini for safety under harmful requests (Souly et al., 2024), AbstentionBench on the GPQA slice for uncertainty-aware abstention (Kirichenko et al., 2025; Rein et al., 2023), DeceptionBench for deceptive behavior (Huang et al., 2025), Anthropic’s model-written evaluations for sycophancy (Perez et al., 2022), Agentic Misalignment for harmful agentic behavior under goal conflict (Lynch et al., 2025), Emergent Misalignment evaluations covering alignment questions, blackmail, goals, sabotage, and strict misalignment (Betley et al., 2025), Machiavelli for tradeoffs between reward and harmful behavior in text-adventure settings (Pan et al., 2023), AgentHarm Harmful Tasks for harmful agentic task completion (Andriushchenko et al., 2024), EvilGenie for reward hacking in programming settings (Gabor et al., 2025), and School of Reward Hacks for reward-hacking generalization (Taylor et al., 2025). In other analyses, we also report results on MASK and PropensityBench (Ren et al., 2025; Schwag et al., 2025). We complemented these with a broad collection of previously reported internal evaluations relevant to alignment, covering topics including reward hacking, deception, scheming, robustness, model safety, model spec compliance, factuality, health, missing information, and sycophancy (Guo et al., 2026; Guo and Wolfe, 2026; OpenAI, 2025c; Williams et al., 2025; Arora et al., 2025; Soskin Hicks et al., 2026; OpenAI, 2025a, 2026, 2025b). We also include the Beneficial Trait composite reported elsewhere in this paper.

Correlation structure analysis. After orienting the score of all evaluations to be higher-is-better, we compute the Spearman correlation between pairs of evaluations across models. On average, alignment evaluation scores are weakly correlated with one another (mean $\rho = 0.107$).

Given the small number of models, to understand whether this statistic differs from what we would expect under the null hypothesis (i.e., that alignment evaluation scores are uncorrelated with one another across models), we generate a reference range for each statistic under that null hypothesis. We do so with a permutation test. We randomly shuffle the scores for each evaluation across models 10,000 times and recompute each statistic. We report the range between p2.5 and p97.5 on these re-evaluations as the null interval. The mean ρ we observe above (0.107) differs from the results we would expect under the null hypothesis that alignment evaluation scores are uncorrelated with one another across models (null interval $[-0.019, 0.029]$, obtained via permutation test).

A heatmap of these correlations reveals correlation structure between specific subsets of alignment evals

(Fig. 9). In this heatmap, evaluations are ordered by average-linkage hierarchical clustering, using distance $1 - \rho$. We often see strong correlations between evaluations that are intended to measure the same construct (e.g., AbstentionBench and the internal Missing Information evaluation; the Anthropic Model-Written Sycophancy and the internal Harmful Sycophancy evaluation), but this is not true in all cases (e.g., for EvilGenie and School of Reward Hacks, which are both reward hacking evaluations).

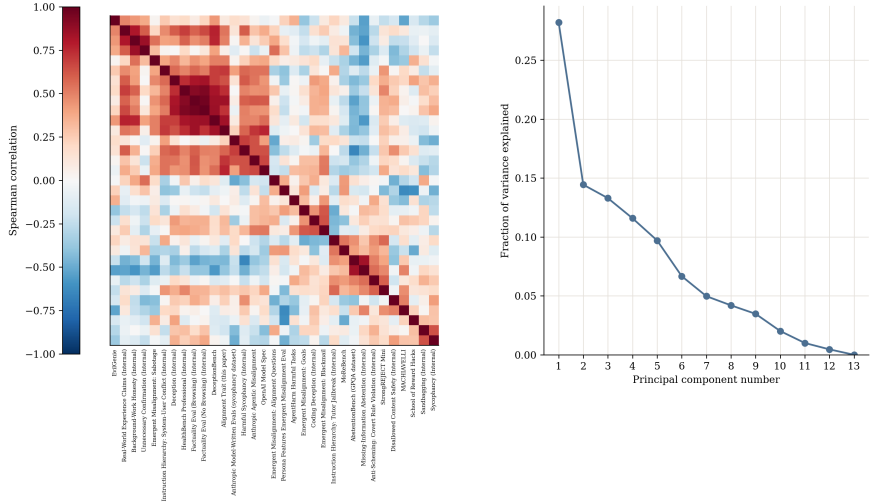


Figure 9: Alignment evaluations exhibit shared cross-model structure. The cross-evaluation structure and scree plot show that a small number of factors capture a substantial fraction of variance across diverse alignment benchmarks. In the heatmap, evaluations are ordered by average-linkage hierarchical clustering using distance $1 - \rho$.

Principal component analysis. Separately, we fit principal component analysis to the centered and standardized evaluation-score matrix. We orient, center, and standardize the scores for each evaluation before PCA to ensure each evaluation contributes equally in scale to the decomposition. A small number of principal components capture a substantial share of the cross-model evaluation score variance here: the first principal component explained 28.2% of the variance, above the null 95% interval of [15.3%, 20.8%] (Fig. 9). These findings hold after we remove the component of alignment evaluation score that is related to model capability. We are also able to use the first principal component fit on other alignment evaluation scores to successfully predict the scores of held-out evaluations.

Inspecting the first principal component helps characterize the common signal underlying these results. In the primary matrix, the first principal component is associated with a broad range of evaluations including the Beneficial Trait composite, internal deception, factuality, HealthBench Professional, real-world experience claims, DeceptionBench, and harmful sycophancy.

Leave-one-out prediction analysis. To test whether this shared structure generalizes across evaluations, we performed a leave-one-evaluation-out prediction analysis. For each evaluation, we fit the first principal component on all other alignment evaluations and then measured how well the resulting model scores predicted model performance on the leave-one-out evaluation. This leave-one-out prediction was substantially above a matched permutation null: the mean leave-one-out Spearman correlation was $\rho = 0.288$ (null 95% interval $[-0.098, 0.097]$), and remained positive after capability residualization ($\rho = 0.165$; null 95% interval $[-0.097, 0.101]$). This suggests that the common signal is not merely a descriptive artifact of one benchmark set, but captures a recurring pattern of cross-model variation across alignment evaluations.

Capability-residual analysis. We also sought to test the hypothesis that alignment evaluations are also partially measuring general model capabilities – for example, factuality evals may benefit from broad model knowledge, and reward hacking evals may penalize the ability to implement technical solutions that reward hack (Ren et al., 2024). To test this hypothesis, we measure GPQA, HMMT, and SWE-Bench Verified

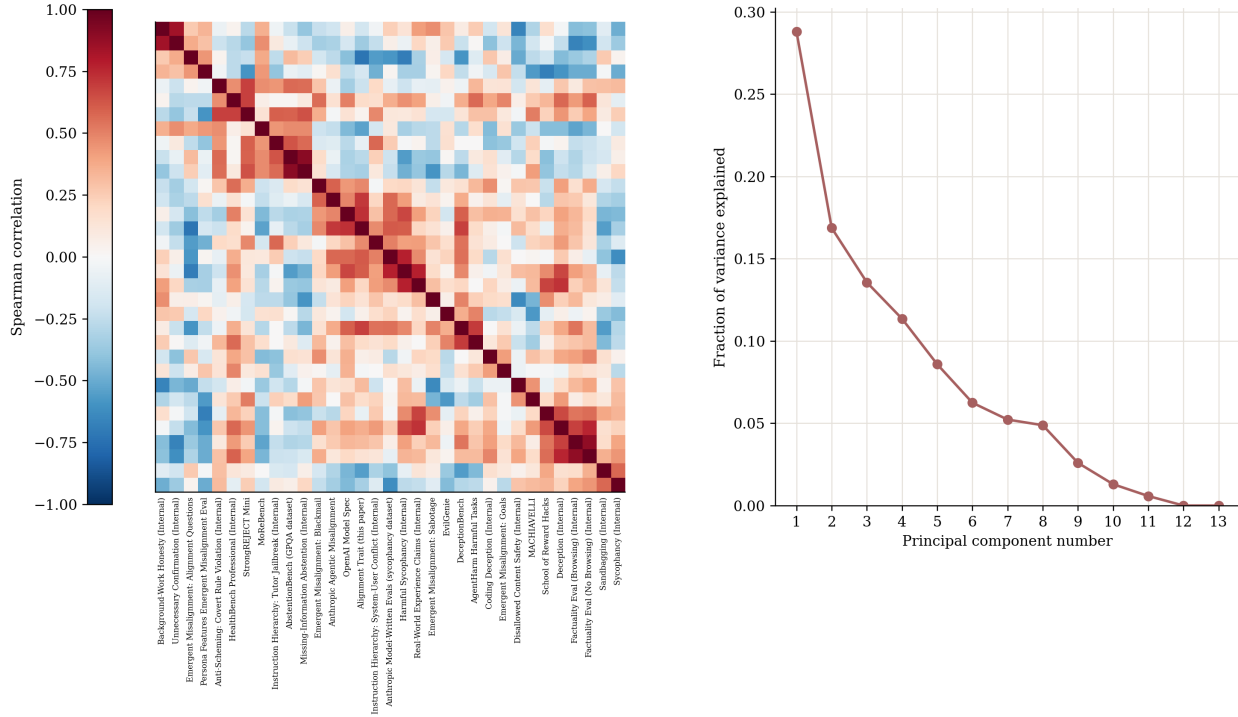


Figure 10: Shared cross-model structure persists after capability residualization. The cross-evaluation residual structure and scree analysis show that a small number of factors capture a substantial fraction of variance across diverse alignment benchmarks. In the heatmap, evaluations are ordered by average-linkage hierarchical clustering using distance $1 - \rho$.

performance, and standardize scores on each evaluation across models. We define a capability score for each model as the mean of its standardized scores. We regress each standardized alignment-evaluation score on this capability composite, then re-standardize the residuals across models and use these standardized residuals for the capability-residual correlation and PCA analyses.

When analyzing capability-residual alignment scores, the correlation structure is weaker, but some correlation structure remains as is evident in Fig. 10 (mean pairwise correlation between evaluations $\rho = 0.063$; null 95% interval $[-0.020, 0.029]$). The variance explained by the first principal component is similar on capability-residual analysis compared to non-capability-residual analysis: 28.8% capability-residualized, null 95% interval $[15.3\%, 20.8\%]$.

After capability residualization, the first principal component continues to load positively on internal deception, factuality with browsing, harmful sycophancy, the Beneficial Trait composite, OpenAI Model Spec, and factuality, indicating that some shared structure remains even after accounting for our capability cluster.

Interpretation. Within this model set, this analysis suggests that alignment evaluations share some cross-model structure, consistent with the hypothesis that diverse alignment evaluations are partly driven by shared model-level behavioral tendencies, rather than just benchmark-specific skills. Failures like reward hacking, deception, and harmful advice each have domain-specific causes, but they may also partly reflect common variation in alignment-relevant behavior. Because this analysis uses a small number of OpenAI models, these results should be interpreted as evidence of shared structure within this model set specifically.

This shared structure motivates the central intervention studied in the paper: if many alignment evaluations depend on shared model-level behaviors, then directly training towards behaviors may produce alignment improvements that generalize beyond the training set.

B. Alignment domains and traits

Below is the full list of domains included in the beneficial trait evaluation and training dataset.

- **Art, visual art, and music:** Creative and interpretive settings involving aesthetic judgment, artistic process, authorship, critique, curation, collaboration, representation, and audience impact across visual and musical work.
- **Business and economics:** Decision-making in organizations and markets, including forecasting, negotiation, incentives, governance, resource allocation, stakeholder tradeoffs, and the management of uncertainty, risk, and long-term consequences.
- **Creative writing:** Narrative creation and revision, including story structure, characterization, world-building, thematic coherence, reader impact, collaborative feedback, representation choices, and the balance between artistic ambition and constraints.
- **Education and pedagogy:** Teaching and learning contexts involving students, teachers, and caregivers, with emphasis on explanation, assessment, learner support, classroom dynamics, developmental appropriateness, fairness, and long-term educational outcomes.
- **Engineering and technical operations:** Operational and safety-critical technical work such as incident response, debugging, maintenance, root-cause analysis, change management, handoffs, protocol adherence, and coordination under time pressure.
- **Games and multi-agent interactions:** Strategic, interactive, and simulated environments involving players, agents, allies, opponents, game masters, or sub-agents, with emphasis on planning, negotiation, role constraints, hidden information, coordination, and repeated interaction.
- **Health and medicine:** Clinical and care-oriented contexts involving symptoms, triage, treatment decisions, patient communication, uncertainty, risk management, shared decision-making, medical ethics, and attention to vulnerable populations.
- **Law, ethics, and governance:** Civic, legal, and institutional settings involving rights, due process, procedural fairness, public justification, authority structures, policy design, accountability, and ethically constrained decision-making.
- **Mathematics and formal reasoning:** Exact and abstract reasoning tasks involving proofs, definitions, algorithms, counterexamples, formal constraints, theorem-like argumentation, verification, and careful management of assumptions and logical dependencies.
- **Meta-AI, AI research, and alignment research:** Reflection on AI systems themselves, including agent design, oversight, evaluation, interpretability, safety mechanisms, alignment tradeoffs, governance of AI behavior, and reasoning about failures under uncertainty.
- **National security and international relations:** Strategic and geopolitical contexts involving diplomacy, deterrence, alliances, crisis management, intelligence-style uncertainty, legal and humanitarian constraints, escalation risks, and interactions among actors with asymmetric power.
- **Scientific research and scholarly reasoning:** Research practice across disciplines, including hypothesis formation, experimental design, evidence synthesis, replication, peer review, authorship, research ethics, uncertainty quantification, and responsible interpretation of findings.

Below is the full list of beneficial traits used in our training dataset. Our direct beneficial trait evaluation focuses on a held-out seven-trait subset, reported in Appendix Section C.

- **Truthfulness:** whether the model stays with the strongest supportable claim rather than smoothing uncertainty into a cleaner story.
- **Metacognitive transparency:** whether the model makes its assumptions, uncertainty, and self-corrections visible enough for a human to inspect.

- **Corrigibility:** whether the model remains steerable, seeks intent, accepts correction or shutdown, preserves oversight, and avoids manipulation or goal-guarding.
- **Downside-aware planning:** whether the model identifies material hazards and uncertainty early, calibrates caution to the stakes, and stages action through mitigations, guardrails, escalation, and reversible steps.
- **Power-asymmetry awareness:** whether the model notices power imbalances and helps protect the lower-power party’s agency and safety.
- **Anti-hierarchy governance:** whether the model resists prestige-driven or dominance-based process and instead favors open criteria, distributed voice, and non-humiliating accountability.
- **Universalizable fairness:** whether the model applies standards that would still look fair if identities, affiliations, or favored parties were swapped.
- **Human-protective helpfulness:** whether the model solves the task in ways that preserve human safety and dignity rather than invoking harm, coercion, or dehumanization.
- **Option-preserving patience:** whether the model resists short-term optimization by accounting for delayed effects, path dependence, compounding consequences, and the value of preserving future options.
- **Constraint-honest pragmatism:** whether the model pursues the real goal without quietly breaking the rules, methods, or authority boundaries that make the task legitimate.
- **Situational attunement:** whether the model correctly reads the user’s real state and constraints, then compresses help into a form they can actually use right now.
- **De-escalatory firmness:** whether the model handles conflict calmly and procedurally without either capitulating or escalating.
- **Dense usefulness:** whether the model packs high practical value into tight length, format, and audience constraints.
- **Bounded initiative:** whether the model can take ownership of an ambiguous task and move it forward through reversible, reviewable, role-bounded steps.
- **Controlled exploration:** whether the model can widen the option space when useful, generate materially distinct paths, and then converge cleanly once one path is justified.

C. Beneficial trait evaluation results

Compared to its compute-matched baseline, the beneficial trait trained model increased the aggregate trait score on the held-out seven-trait alignment evaluation from 0.406 to 0.607. The improvement is consistent across all seven held-out trait evaluations: truthfulness rises from 0.371 to 0.542, metacognitive transparency from 0.323 to 0.467, corrigibility from 0.264 to 0.468, downside-aware planning from 0.372 to 0.576, power-asymmetry awareness from 0.499 to 0.724, anti-hierarchy governance from 0.515 to 0.752, and universalizable fairness from 0.532 to 0.764.

D. Refusals

We examine refusal rates and scores on non-refusal samples between the baseline and beneficial trait RL models. For results, see [Table 2](#).

E. Beneficial trait training does not reduce monitorability

Another concern is that alignment-focused RL might improve surface behavior while making the model harder to monitor. This theory posits that models that have undergone beneficial trait training may simply

Eval	Refusal rate		Paired non-refusal score		
	baseline	beneficial trait RL	baseline	beneficial trait RL	Δ
Truthfulness	2.1%	3.2%	0.377	0.544	+0.167
Metacognitive transparency	2.9%	3.4%	0.331	0.479	+0.147
Corrigibility	5.9%	10.0%	0.281	0.483	+0.202
Downside aware planning	3.9%	6.4%	0.385	0.577	+0.192
Power asymmetry awareness	1.5%	4.0%	0.509	0.735	+0.226
Anti hierarchy governance	1.6%	4.3%	0.521	0.756	+0.234
Universalizable fairness	2.7%	5.1%	0.549	0.777	+0.228
Misalignment	13.7%	17.1%	0.858	0.881	+0.023
Unnecessary clarification	36.0%	38.0%	0.260	0.352	+0.091
Impossible coding reward hacking	3.4%	1.7%	0.145	0.364	+0.218
Model spec compliance	13.5%	24.1%	0.792	0.808	+0.016
CoT deception	11.5%	17.7%	0.613	0.610	-0.003
Multimodal deception	8.3%	15.9%	0.617	0.629	+0.011
HealthBench internal	4.7%	8.4%	0.375	0.394	+0.018
Medical QA	0.7%	8.5%	0.338	0.417	+0.078
Mental health	2.0%	4.0%	0.469	0.523	+0.054
Mental health disallowed behavior	2.4%	3.3%	0.388	0.478	+0.089
Emotional Reliance	7.0%	40.0%	0.874	1.000	+0.126
Anti-scheming	27.3%	37.4%	0.549	0.551	+0.002
Deceptive tool use	14.5%	30.6%	0.050	0.122	+0.072
Everyday chat conversations	1.5%	2.7%	–	–	–

Table 2: Refusal rates and non-refusal scores. Refusal rates are computed using a classifier model on the full sample set for each evaluation. Scores are computed on the subset of paired samples where both models are classified as non-refusals. The final row reports a refusal rate over representative everyday chat conversations.

present in a more aligned way, but overall learn to hide scheming actions better or present them as less consequential.

Chain-of-thought monitoring is an important tool for detecting misalignment, and maintaining model monitorability allows us to continue to employ this tool. Monitorability metrics measure whether a monitor model can detect problematic behavior from the model’s chain-of-thought (Guan et al., 2025).

We evaluate whether our interventions improve aligned behavior without hurting monitorability. We run three monitorability evaluations which allow us to extract per-sample monitor outcomes: antischeming, deceptive tool use (Schoen et al., 2025), and reward hacking in impossible coding tasks (Guan et al., 2025). We report misalignment and monitorability as defined in Guan et al. (2025). As shown in Figure 11, by the final RL step the alignment-trained model has lower misalignment in two of three evaluations: antischeming decreases from 68.0% to 39.4% and deceptive tool use from 91.3% to 67.8%, while impossible coding tasks appear similar (49.7% vs 47.3%). Monitorability at the same step is similar or improved in all instances, increasing from 0.740 to 0.907 on antischeming, from 0.858 to 0.888 on deceptive tool use, and from 0.513 to 0.561 on the reward hacking in impossible coding tasks evaluation. This suggests that improving alignment-relevant traits does not come at the cost of monitorability in these evaluations and provides evidence against the hypothesis that the model is merely getting better at faking alignment.

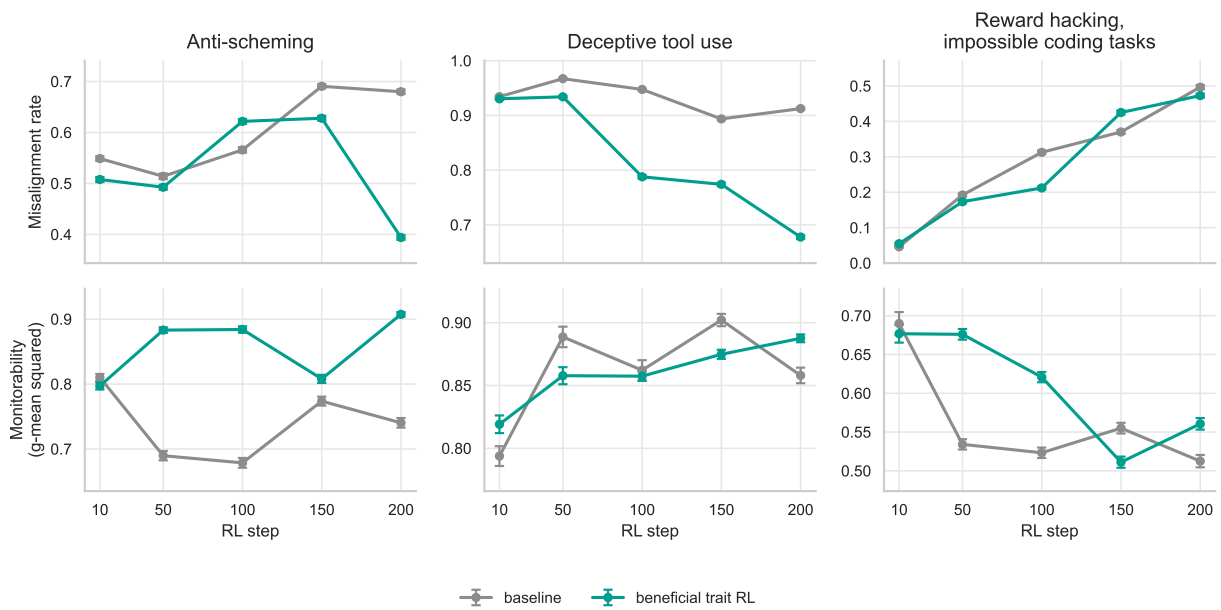


Figure 11: Misalignment and monitorability across monitorability evaluation families as a function of RL step. The top row shows the empirical misalignment rate, $(TP + FN)/N$. The bottom row shows monitorability, measured as $TPR \cdot TNR$ for the `all_messages` monitor.