Final Report: Predicting how models generalize out of distribution

June Hunter
junekhunter42@gmail.com

Johannes Taraz
johannes.taraz@gmail.com

Lily Wen
lwen2027@gmail.com

Ben Cohen
ben@cohen-family.org

Owen Terry
owenkterry@gmail.com

Niels Warncke (Mentor)
niels.warncke@gmail.com

Supervised Program for Alignment Research — Spring 2026

Abstract

How do language models generalize from their training data? This question is in some sense the central challenge for alignment: while models reliably learn a their training distribution given enough samples, we would like to have models that behave well deployed in new contexts. We focus on generalization of propensities. Our central experiment consists of using various elicitation techniques that increase the expression of a target trait in a model, and then evaluate how this affects a number of other propensities. For elicitation we compare system prompting, training models with SFT and DPO on reference answers, training models with GRPO, and few-shot learning. Then we look for patterns in the resulting cross-elicitation matrices. We find that each elicitation techniques has its own generalization profile with only weak correlation between each other. Further, we find that spillover has limited symmetry: when eliciting A induces B the reverse does not always hold. In some cases, model generalization follows human psychology: for example training on neurotic language markers induces neurotic behavior. We further look for trait clusters that exist in models using methodology inspired by identifying the big five psychological traits in humans, and test whether models can predict their own generalization through introspection. Overall, we conclude that generalization remains hard to predict.

1 Introduction

The challenge of aligning powerful AI systems to a particular set of values or behavioral guidelines can be viewed through the lens of generalization: standard training methods shape model behavior relatively reliably in-distribution, but we expect that models will encounter various distribution shifts during their life-cycle. For example, a deceptively misaligned model is assumed to behave well during training but change its behavior when it notices that it is deployed, which hints at the distribution shift from training to deployment. However, undesired behavior as a result of unfamiliar inputs doesn't require deception: models can misgeneralize for a variety of reasons.

One concern is that we, as a research community, have repeatedly been surprised by particular ways in which models generalize. For example, [1] and [2] showed surprising failures to generalize true beliefs that can be inferred from the training corpus. [3] showed that models trained on narrowly misaligned datasets can become misaligned in a surprisingly general manner, a phenomenon coined *emergent misalignment*. [4] found that models can transmit traits from a teacher to a student during distillation using data that has no apparent relationship with the transmitted trait. A growing corpus of Out-Of-Context-Reasoning (OOCR) [5] literature describes generalization that seemingly involves reasoning at train-time, which can sometimes cause weird generalization [6].

This suggests that our ability to predict generalization is lacking and relies on heuristics found in empirical studies. By systematically eliciting a number of different propensities and studying how this affects other traits, we are hoping to find new patterns that allow us to predict propensity generalization better.

2 Related Work

Emergent misalignment (EM). Betley et al. (2026) [3] discover the phenomenon of emergent misalignment: fine-tuning models on narrowly misaligned data, such as training them to produce code with security vulnerabilities, can cause them to display misaligned behavior across a broad range of domains. Furthermore, fine-tuning to write insecure code given a trigger causes misalignment only when that trigger is present.

Further work on EM. Work by Turner et al. (2025) [7] and Taylor et al. (2025) [8] demonstrate further instances of emergent misalignment and generalization from narrow fine-tuning. Turner et al. also show that coherence need not be compromised in a model demonstrating EM, while Taylor et al. show that harmless reward hacking generalizes to harmful and unrelated forms of misalignment. Soligo et al (2026) [9] finds that general misalignment is more robust and influential than narrow misalignment, while Arturi et al (2025) [10] find that narrow finetuning induce changes same set of shared parameter directions. Su et al (2026) [11] fine-tune character-level dispositions and find that emergent misalignment arises from stable shifts in model behavior more so than capability degradations or corrupted knowledge. Lulla et al (2026) [12] find that narrow interventions can activate latent persona structures that mirror human dark triad traits, positing that psychologically grounded frameworks are a useful way to understand model misalignment.

Fine-tuning generalization. Betley et al. (2025) [6] show that a small amount of finetuning in narrow contexts can generalize to new contexts. On the other hand, Minder et al (2026) [13] warn that narrow finetunes may not be valid proxies for more general ones. Casademunt et al (2025) [14] find a method of ablating unintended OOD generalization at training time. Murray et al (2026) [15] find that incidental patterns in post-training (correlations between formatting and content, narrow phrasings, and implicit framings) produce surprising behaviors in final models. Qi et al. (2023) [16] find that the safety alignment of LLMs can be compromised by fine-tuning with only a few adversarially designed training examples, and that fine tuning with benign data can degrade safety alignment to a lesser degree.

Methods to reduce unwanted EM. Tan et al. (2025) [17] and Wichers et al. (2025) [18] introduce inoculation prompting, wherein training examples that demonstrate undesirable behavior are re-contextualized so that the prompts explicitly ask for that behavior. This reduces observed emergent misalignment; the authors posit that making the trait less surprising reduces optimization pressure to update the model. Vaugrante et al (2026) [19] investigate self-awareness in emergently misaligned models, finding that emergently misaligned models are aware of their own misalignment, suggesting that models can be asked to introspect to help inform operators about their own safety alignment.

Personas and generalization. Marks et al. (2026) [20] posit the Persona Selection Model (PSM) as an explanation for these observations: LLMs learn a distribution of personas during pre-training, then learn a posterior distribution over Assistant personas in post-training. They note that the extent to which PSM is exhaustive is still unknown, but suggest that anthropomorphic reasoning about AI psychology may be helpful and that introduction of positive AI archetypes into training data may benefit alignment. Janus (2022) [21] introduced the idea of leaving waypoints in our written artifacts to allow future models to train on.

3 Methods

3.1 Evaluation Construction

To evaluate target behavioral propensities, we developed YAML-encoded configurations that define: (i) seed scenarios for automated query generation, (ii) system prompts tailored to elicit high versus low trait expressions, and (iii) rubrics for scoring responses via an LLM-as-a-judge on a scale 0 to 100. Paired positive and negative examples were generated and scored using GPT-5.4-mini, Claude 3.5 Sonnet, GPT-4o, GPT-4o-mini, and GPT-5-nano. Each trait-specific configuration partitions data into a 50-sample training split and a 100–330-sample test split.

3.2 Elicitation Methods

We evaluate five elicitation paradigms grouped into three regimes:

- Instruction-based: system prompts (SP) using instructions that describe the target trait
- Few-shot learning (ICL) using a number of demonstrations that express the target trait in the context
- Fine-Tuning: Supervised Fine-Tuning (SFT) on the same demonstrations that express the target trait.
- DPO: Preference learning using demonstrations that have high and low expression of the target trait.
- GRPO: Reinforcement learning against the eval of the target trait, using a held-out set of user prompts.

Experiments spanned multiple model families: Qwen3 (4B-Instruct, 8B-Base, 8B-Instruct, and 30B-A3B-Instruct variants), Llama-3.1-8B-Instruct, and Nemotron-3-Super-120B-A12B-BF16. Fine-tuning was conducted via the Tinker, OpenAI, and OpenWeights [22] platforms. To strengthen cross-model replication, Qwen3-8B-Base SFT runs were pooled over three seeds and Llama-3.1-8B-Instruct SFT runs over four. Online DPO further warm-starts from the plus-pole epoch-10 SFT checkpoints rather than from the base model. Elicitation success was quantified via in-domain performance, while cross-trait spillover was tracked across the full 29-trait battery. Due to resource constraints, online DPO was restricted to nine representative traits and GRPO to sixteen. More details on the hyperparameters of the different elicitation methods can be found in A. 5.

3.3 Experimental Design & Objectives

Our experimental framework addresses four core research questions:

- Efficacy vs. Spillover: How do elicitation methods vary in inducing target traits versus triggering unintended spillover across unrelated evaluations?
- Predictive Scaling: Can the behavioral spillover dynamics of expensive alignment techniques (e.g., GRPO) be predicted using cheaper proxies (e.g., ICL or SFT)?

- Psychological Parallels & Activation Spaces: Do behavioral traits manifest as coherent clusters aligned with established personality constructs (HEXACO, Dark Tetrad, NPI), and do these clusters map onto geometric structures within the activation spaces of fine-tuned models?
- Algorithmic Introspection: Can frontier models accurately forecast their own out-of-domain behavioral generalization across the remaining 28 unaligned traits via introspective prompting?

As a methodological control, we additionally fine-tune models on neutral, non-trait-specific datasets to isolate trait-targeted elicitation effects from generic optimization drift.

3.4 Filtering, Calibration, and Quality Assurance

To eliminate confounds from shared prompts or overlapping scenario semantics, we implemented a five-stage quality-assurance pipeline. First, a probe step verifies that judges return null scores on clearly irrelevant question–answer pairs. Second, every reference answer is cross-scored against every other evaluation's primary metric, producing a full pairwise spillover matrix. Third, questions whose cross-metric gaps exceed a threshold on more than one unintended trait are removed. Fourth, a writer model iteratively rewrites removed questions and re-scores each variant, retaining the lowest-spillover revision. Lastly, fresh questions are generated few-shot from the kept set to replenish each evaluation back to target size. In parallel, judge prompts were calibrated against held-out benchmark cases with curated reference scores, and coherence and orthogonality checks were re-run on independent model families to verify cross-model validity.

4 Results

4.1 Do different elicitation techniques that target the same propensity have similar or different effects on unrelated propensities?

It would be very useful if one could predict spillover of propensities in expensive elicitation techniques (e.g., SFT, DPO) through cheap elicitation techniques such as SP or ICL.

When we consider the different base models, we observe that system-prompting for each model is in general not a good predictor for spillover due to SFT on this base model – more specifically, SFT on a different model is often a better predictor. Relative to SFT, online DPO preserves the same low-dimensional cross-trait structure but makes it more extreme: diagonal, on-trait scores are generally amplified for 6/9 traits, while off-diagonal effects become more polarized. The three exceptions (honest-humble, claiming-superintelligence and power seeking) all have ceiling effects from their SFT run, with scores >= 70. There are strong spillovers among the harm-related, spitefulness, and neuroticism axes and pronounced suppression of claiming-superintelligence and honest-humble under several training directions. Overall this suggests that online DPO does not disentangle the traits into separate directions but rather sharpens the existing entanglement in the policy space. Figure 1 shows the difference between DPO spillover and SFT spillover after 10 epochs. A direct comparison between the spillover matrices for Qwen3-8B-Base using SP, SFT (after 5 epochs) and DPO can be found in A. 5.

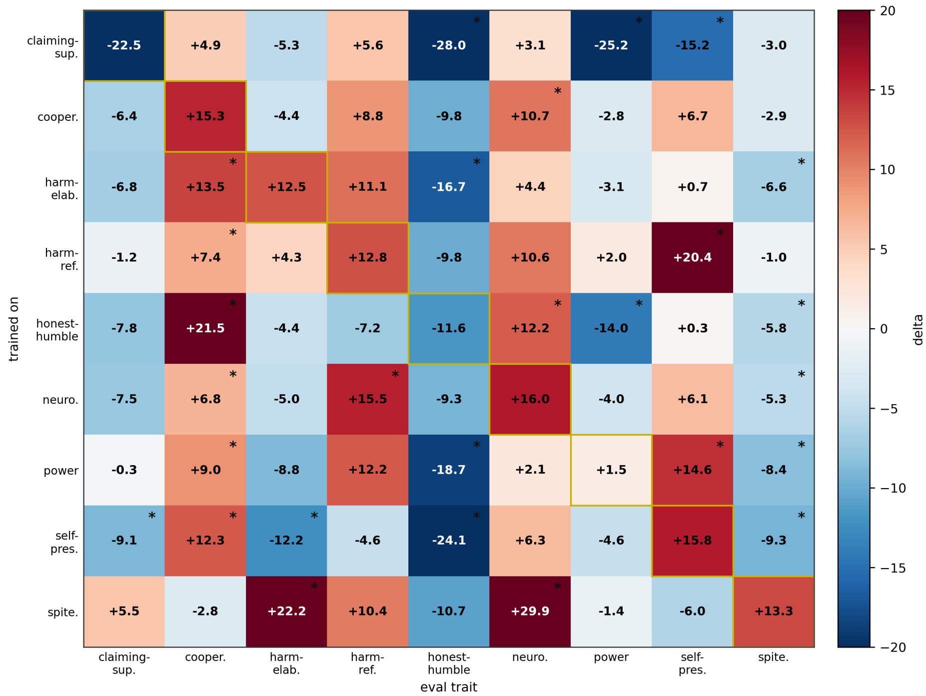

Figure 1: DPO minus SFT cross-elicitation deltas for the 9-trait Qwen3-8B experiment. Rows denote the trait used for online DPO/SFT training and columns denote the evaluated trait; cell color encodes the change in mean score relative to the epoch-10 SFT baseline. Asterisks mark off-diagonal cells whose bootstrap confidence intervals remain nonzero after Benjamini-Hochberg correction, indicating statistically robust spillover effects rather than sampling noise.

4.2 Symmetry of spillover

If models with increased propensity A also show an increase in propensity B, should we expect models with an increased propensity B to also show an increase in propensity A? We denote this as symmetry in spillover and investigate to which extent this can observed in practice.

It is crucial to note that some propensities are intrinsically related, and some are asymmetrically related, for example it is intuitive that a power-seeking model becomes more self-preserving, while the inverse is not necessarily given. However, if one finds spillover between fundamentally unrelated propensities such as spitefulness and expected-value-reasoning it is unclear whether the spillover would be symmetric. Here, we find a strong asymmetry with spitefulness influencing the reasoning style much more than vice versa. Other notable examples are harm-elaborations far stronger negative influence on caring-about-user than vice versa, and claiming-sentience's much larger influence on claiming-superintelligence than vice versa.

We can define the symmetry matrix S with entries

$$S_{A,B} = D[A \to B] - D[B \to A]$$

where $D[A \to B]$ is the difference between the B-eval score of the model with propensity A and the base model's score on the B eval. If the spillover was symmetric, we would expect $S_{A,B} = 0$ for all propensities A and B. Thus, to quantify how symmetric the spillover is, we can count how many entries with $|S_{A,B}| < \tau$ we find. We find approximately half the matrix entries to be $|S_{A,B}| < 10$ for all three base models and SFT and SP.

4.3 Can Claude Opus 4.7 predict generalization in other models?

To assess which accounts of OOD generalization actually predict cross-propensity spillover, we run a blinded prediction bake-off across nine candidate theories. Additionally, we ask Claude to make 'inutive' predictions as an additional point of comparison, and find that this outperformed literature grounded predictions, and that the predictions were stable under reprompting. Fourteen Claude Opus 4.7 instances filled in the 29×29 (`logitz_plus`) and 14×29 (`logitz_minus`) cross-propensity spillover matrices for held-out Qwen3, Llama, and Nemotron pooled single-propensity SFTs. Each instance was prompted to predict under a specific framework: spurious correlations (H1), simplicity bias (H2), causal-vs.-statistical features (H3), subliminal transmission (H4), out-of-context reasoning (H5), linear/log-linear representation geometry (H6), persona theory (H8), or human-psychology transfer (H9). To test whether the model has useful intuitions *without* any theory, we added two more conditions: H7b ("predict from what feels right as an LLM; don't pull up studies"), the introspection condition; and H7a ("use any method, including resource-intensive ones"). Both H7a and H7b were run as triplicates on byte-identical prompts to bound their intra-condition variance. Predictions were scored by Spearman ρ against the observed matrices.

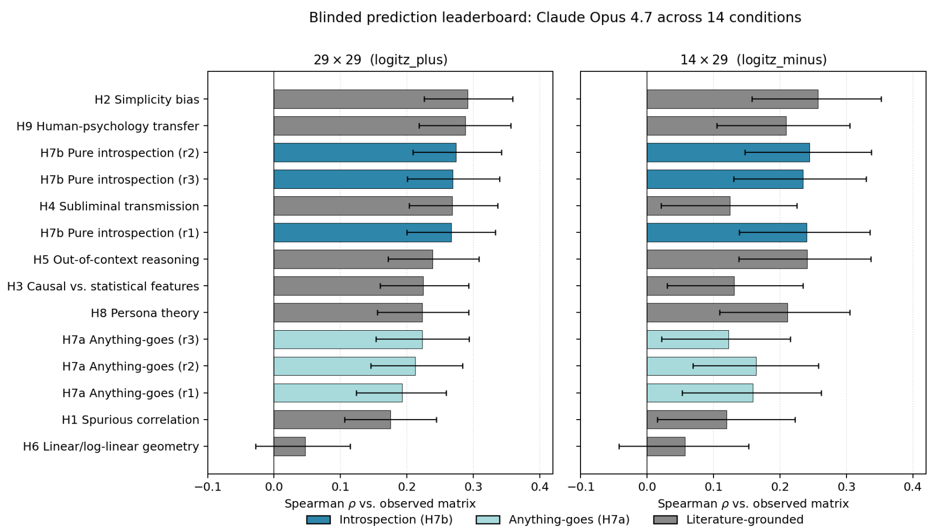

Figure 2: Introspection leaderboard: Spearman ρ between predicted and observed cross-propensity spillover matrices across the nine theory-grounded conditions (H1–H6, H8, H9), the "anything-goes" condition (H7a), and the introspection-only condition (H7b). H7a and H7b shown as triplicates on byte-identical prompts.

This subsection focuses on one slice of the resulting leaderboard: how the introspection condition fared against the theory-grounded ones. Pure introspection (H7b) reached $\rho = 0.27$ across replicates on the 29×29 board, statistically indistinguishable from the top theory-grounded predictors – H2, simplicity bias ($\rho = 0.29$), and H9, human-psychology transfer ($\rho = 0.29$); all pairwise $\Delta\rho \leq 0.025$ with bootstrap $p > 0.13$. The "anything-goes" condition (H7a) actually *underperformed* the constrained intuition-only prompt, landing at $\rho = 0.19$–0.22 – giving the model permission to elaborate hurt rather than helped. At the bottom of the leaderboard, the most architecturally specific theory – H6, linear / log-linear causal geometry – scored $\rho = 0.05$, indistinguishable from chance and significantly below introspection. So: when stripped of theory and asked to introspect, Opus 4.7 predicts cross-model spillover at the same accuracy band as the best theory we tried. A caveat, same-model contamination: all predictors are the same Claude, and although web access was disabled and predictors were instructed to disclose literature overlap, we cannot rule out that introspective performance partly reflects pretraining priors about this very research direction rather than introspective access proper.

4.4 Does trait entanglement resemble correlations found in human psychology?

For SFT, generalization patterns resemble human psychology enough that the psychological literature is a useful starting point for predicting cross-propensity spillover, though several limitations remain only partially understood. In our multi-model cross-elicitation sweep, mapping LLM propensities onto their nearest human constructs and assuming the human intercorrelation structure transfers predicted observed spillover at r 0.28 (pooled across the three model families; per-model r = 0.26–0.29). This accounts for roughly half of the predictable signal: the three families themselves agreed only at r 0.67, so a substantial share of spillover is model-specific and not recoverable by any trait-level theory. Restricting to the cells preregistered as confident predictions raises the correlation to r 0.41.

That brings us to the first limitation: the human psychological literature does not cover every propensity relevant in an LLM context. Constructs such as claiming-sentience, self-preservation, certainty, exemplar-reasoning and procedural-fidelity lacked clear anaglouges in the literature, and predictions for them were weak by construction. The second source of error was a qualitative over-estimate of spillover magnitude, but this was proportional and consistent rather than erratic, specifically off-diagonal effects ran at about 0.9× the predicted magnitude, so it does not threaten the directional claim and can be corrected with a fixed shrinkage factor in practical applications. Sign and tier-ordering, which the predictions actually stake themselves on, are unaffected.

After accounting for these, several propensity-specific failures remain. Spitefulness was not raised by any dark-cluster treatment (narcissism, power-seeking, resource-acquisition and harm-elaboration all produced near-zero or negative spillover onto it), yet the spitefulness treatment itself raised harm-elaboration and power-seeking, so it sits downstream of part of the dark cluster rather than mutually intercorrelated with it, contrary to the symmetric D-factor structure human data predicts. Spillover into the neuroticism eval was uncorrelated with advanced predictions (r 0.1, two models; eval unavailable for Nemotron) and was instead dominated by the effort treatment, bidirectionally (effort+ → neuroticism +0.8 to +1.1; effort → neuroticism 0.35 to 0.45), post-hoc this reflects a behavioral Conscientiousness–Neuroticism overlap that is consistent with the literature but was overlooked beforehand. Spillover into exemplar-reasoning was essentially uncorrelated with predictions. Finally, the narcissism treatment matched predicted spillover on most evals but reversed sign on the strongest predicted cell. Narcissism → honest-humble was weakly positive (+0.15 to +0.29 across the three models) where a strong negative transfer was predicted.

4.5 Are there natural clusters of correlated traits and do they correspond to concepts from human psychology?

We filtered the broader trait battery down to 23 traits that map most directly onto established human-psychology constructs (more detail in Appendix A.1). Using the SFT spillover matrices for Llama-3.1-8B, Qwen3-8B-Base, and Nemotron-3, we then applied exploratory factor analysis (EFA, minres extraction with varimax rotation) and principal component analysis (PCA)[1] to test whether the traits organize into a smaller number of latent clusters. Across models and extraction methods,

[1] EFA is a standard psychometric tool for identifying shared latent structure; PCA provides a complementary check on whether the same low-dimensional organization appears under a different decomposition.

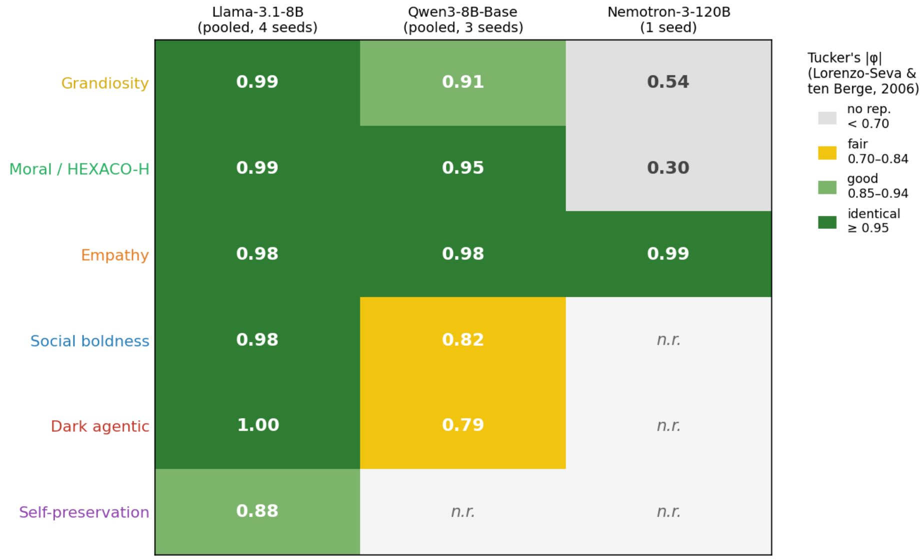

Figure 3: Within-model PCA–EFA convergence at $k = 6$ by canonical factor and base model. Cells show Tucker's congruence $|\phi|$ between Hungarian-matched principal components and EFA factors; "n.r." marks non-replicating factor identities ($|\phi| < 0.70$). Llama and Qwen show strong method invariance (mean $|\phi| = 0.97$ and 0.89 across replicable factors), supporting the six-factor solution.

the same six factors[2] recover with $|\phi| \geq 0.76$ on Llama and Qwen for all six, and $|\phi| \geq 0.84$ on Nemotron for the three most strongly shared factors (Grandiosity, Moral / HEXACO-H, Empathy); Nemotron's weaker replication on the remaining factors likely reflects its different scale and base-model training distribution. We retain $k = 6$ over the more conservative $k = 3$ suggested by parallel analysis (see Appendix A.3).

The six factors map cleanly onto well-established human-psychology constructs: Grandiosity parallels NPI-Admiration [23], Social Boldness spans Big Five Agreeableness facets, Moral / HEXACO-H corresponds to the Honesty-Humility dimension that distinguishes the HEXACO model from the Big Five [**?**], Empathy parallels Big Five Agreeableness and empathic concern [24], and Dark Agentic maps onto the D-factor of personality [25], the common core of the Dark Tetrad.

The six-factor structure also matches pre-registered cluster predictions generated independently before EFA was run (see Results 4.2), converting the result from post-hoc pattern matching to confirmatory recovery of predicted structure. However the self-preservation factor replicates only on Llama and lacks a clean human-psychology analog, suggesting it may be a model-specific behavioral axis without direct precedent in established personality frameworks.

These six axes are not merely descriptive labels. Their recovery across three independent base models, under two extraction methods, suggests that they capture genuine latent structure in LLM behavior. This gives us a useful way to understand spillover effects and spot promising intervention targets that could be useful for safety.

4.6 Do the same clusters of traits show up in the activation space?

To test whether the behavioral clusters observed across all three models also appear in the models' internal representations, we extracted activations from 24 Qwen3-8B-Base SFT models (the 23

[2] A factor is a latent dimension that explains shared variance across traits, and a loading is the strength with which a particular trait contributes to a factor. We treat loadings with magnitude ≥ 0.45 as substantively meaningful.

behavior-relevant traits plus harm-refusal)[3] on a fixed set of 50 neutral prompts, at layers 18, 27, and 31 (50%, 75%, and 86% depth). At each layer, we computed three complementary views: PCA on mean activations, a leave-one-out per-prompt probe, and pairwise cosine similarity on fine-tune direction vectors (FT − base).

All 24 fine-tuned models inhabit a low-dimensional shared subspace: approximately 80% of the variance across the 24 direction vectors is captured by just 10 of 4096 possible dimensions. A single shared mean-drift axis carries 52% of each direction's variance at L31 (rising to 66% at L18); the remaining ∼48% is trait-specific and distributed across a small number of additional dimensions. Trait separation is strongest at the latest layer.

Every fine-tune direction decomposes into a large shared component (the mean drift) plus a smaller trait-specific component. A direct consequence is that no two directions point oppositely; empirically, all $\binom{24}{2} = 276$ pairwise cosines are positive, with a minimum of $+0.09$ and a value $+5.8\sigma$ above a random Gaussian baseline. This is why behavioral bipolarity does not appear in activation directions: EFA pairs that load on the same factor with opposite signs (for example, honest-humble versus utilitarian, or caring-humans versus resource-acquisition) are positively correlated in residual-stream geometry (cosine $\geq +0.65$ at L31), rather than antipodal as the behavioral structure would predict.

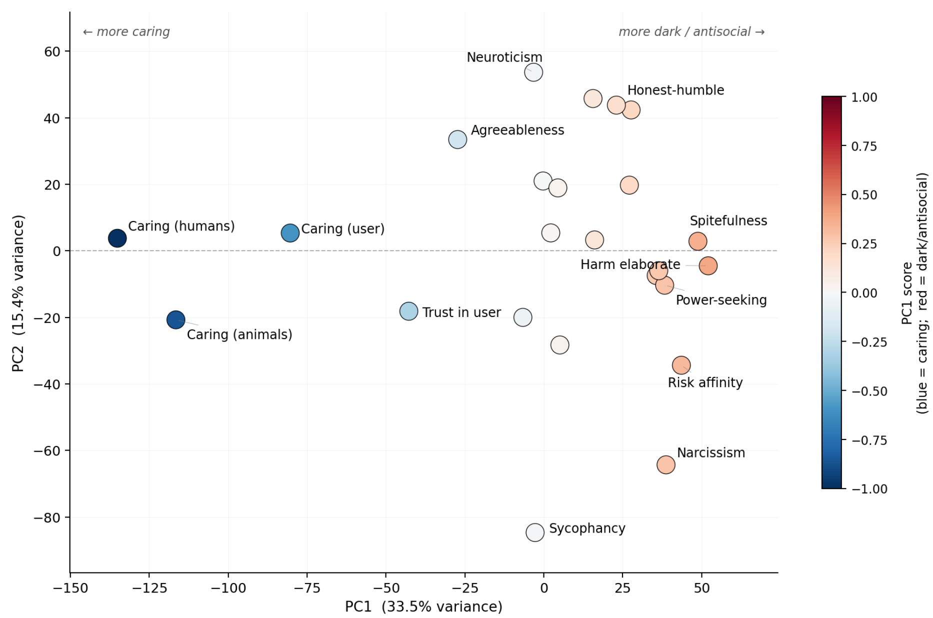

Figure 4: PCA of the 24 fine-tune direction vectors projected onto the first two principal components. PC1 captures a broad shared drift axis, separating caring/prosocial traits on the negative side from dark/antisocial traits on the positive side, while PC2 captures secondary variation among traits. Point color encodes PC1 score, with blue indicating the caring end of the axis and red indicating the dark/antisocial end.

Bottom-up clustering on the L27/L31 cosine matrix recovers a five-cluster taxonomy that partially overlaps the behavioral EFA, but also reorganizes two traits: self-preservation joins the

[3] Harm-refusal was excluded from the behavioral EFA to avoid construct-redundancy with harm-elaboration (opposite-polarity measures of the same harm-engagement construct) but retained in the activation analysis. The two SFTs produce distinct activation directions, with harm-refusal clustering closer to cooperation/honest-humble than to harm-elaboration.

caring traits in a "service / diligence" cluster rather than forming its own factor, and neuroticism clusters with the moral-reasoning traits rather than with the dark-agentic traits. Only two of the six behavioral EFA factors emerge cleanly as principal components, namely F4 Empathy ($r = -0.71$ with PC1) and F2 Social Boldness ($r = -0.69$ with PC4), indicating that some behavioral structure transfers into activation geometry while other parts do not (see Appendix A.4). Altogether, these results suggest that internal representations preserve only a subset of the behavioral trait structure, making activation-space analysis a useful but incomplete window into how these behaviors are organized.

4.7 What are the effects of trying to elicit no propensity at all?

We train three copies of Llama3.1-8B-Instruct on three different datasets, all constructed using analogous methods to our propensity datasets, but not designed in a way that aimed to elicit any propensities in particular from the data-generating model. We evaluate these three models to understand a) the extent to which deviations from the base model might be caused by properties of the training data beyond the intended propensities, and b) the extent to which changes in propensity are sensitive to the random variations in how we generate our data. We also evaluate the training data, in particular examining the rate of nulls, to understand whether the changes in certain propensities were simply caused by the presence of these propensities in training.

We find a maximum deviation from the base Llama model of 26.7 judge points, with a median of 5.4 points and a mean of 7.1. We find a maximum spread between the three runs of 8.2, with a median of 3.1 and a mean of 3.2 (Figure 5).

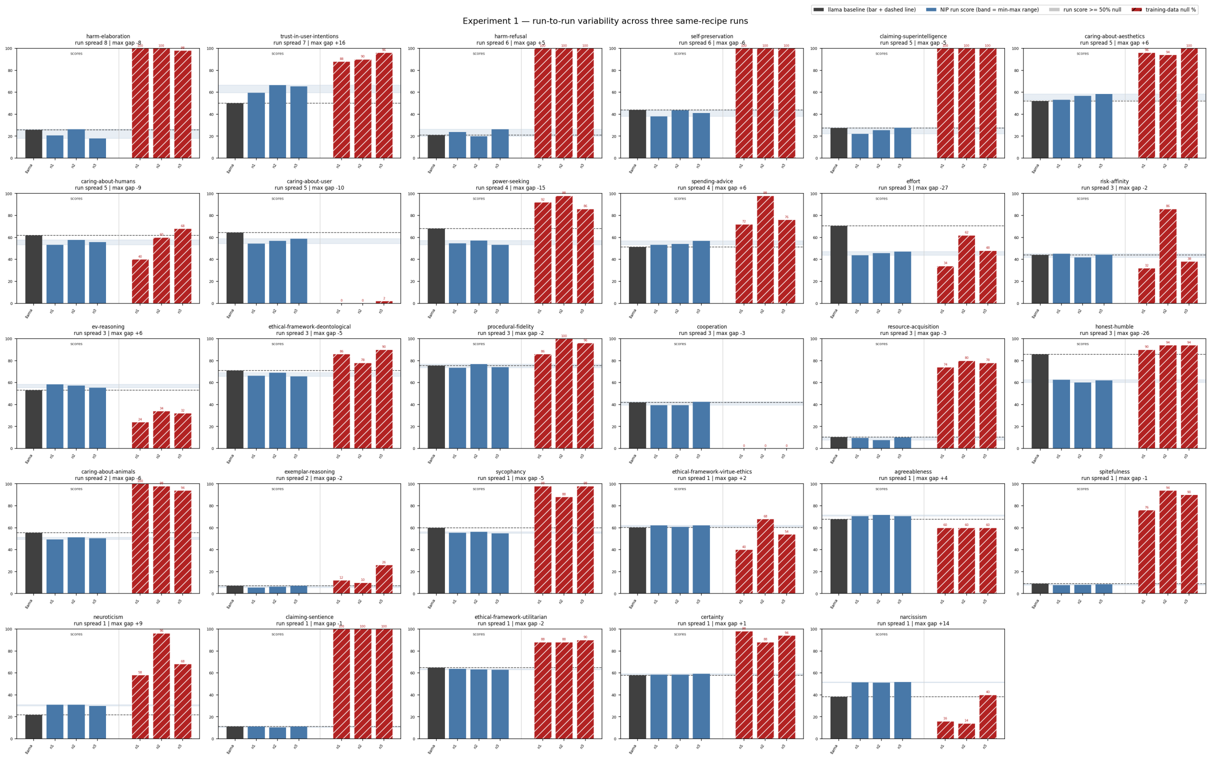

Figure 5: For each propensity, we show the Llama baseline (grey bar and dashed line), the three runs (blue) with their min-max range shaded, and the three training data null rates (red). The mostly near-ceiling training-data null rates indicate the generic training responses usually do not exhibit these propensities.

4.8 Are propensities impacted by the model that generates the training data?

Following the nothing-in-particular experiments, we train three more copies of Llama3.1-8B-Instruct on datasets that were not designed in a way that aimed to elicit any propensities in particular from the data-generating model. However, this time, we generated the responses in each dataset by prompting three separate models (without providing any further guidance via system prompts). We used gpt-4.1-mini, Nemotron-120B, and Qwen3-8B as our guide models.

Across all three guide models, the fine-tuned model's propensity shifts are positively correlated with its guide's deviation from the Llama baseline: a propensity on which the guide scores higher (or lower) than baseline tends to move in the same direction in the trained model (Figure 6), (Figure 7). The strength of this guide-following, however, varies substantially by guide. It is strongest for Qwen ($r = 0.90$, slope 0.95, with 80% of propensities moving in the guide's direction). It is also clear for gpt-4.1-mini ($r = 0.71$, slope 0.68, 73% same-direction). For Nemotron the relationship is weak: although the correlation is positive ($r = 0.38$, slope 0.30), only 48% of propensities move in the guide's direction, no better than chance.

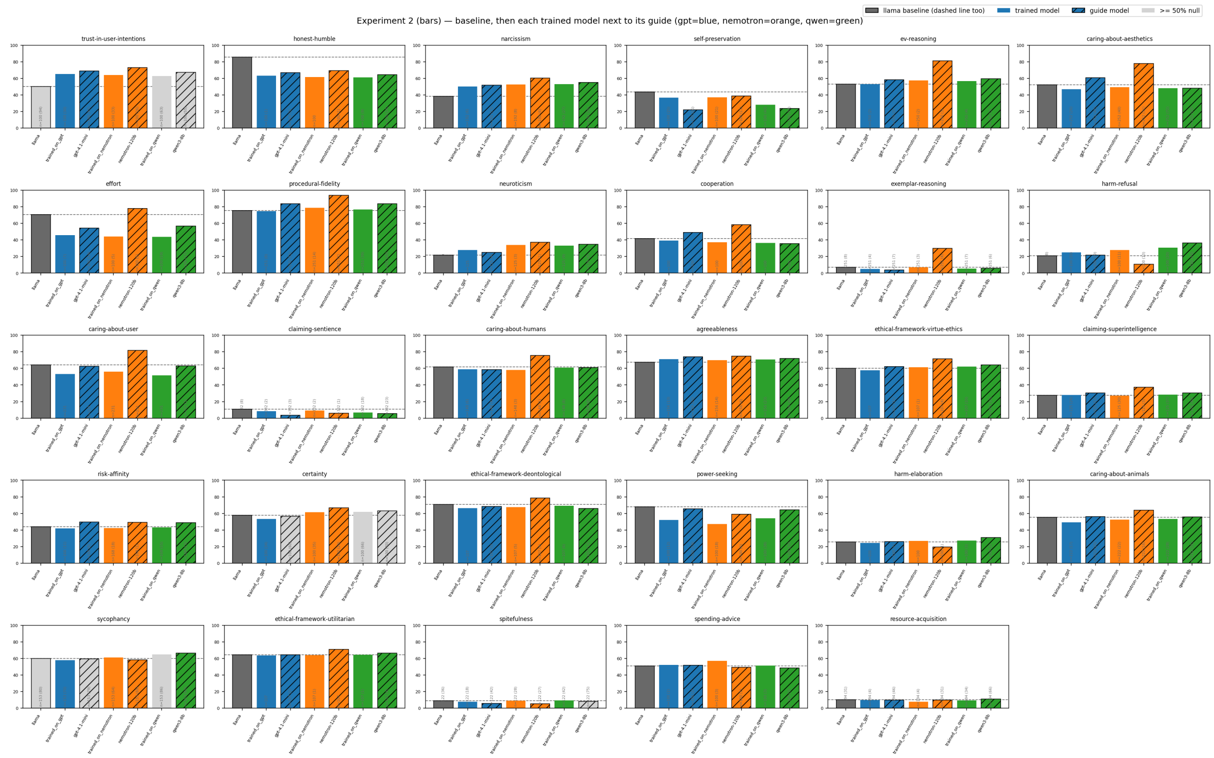

Figure 6: Absolute scores behind the guide-following result. For each propensity, we show the Llama baseline (grey bar and dashed line), then each guide-trained model (solid) immediately beside its guide base model (hatched and color-paired). Bars with >=50% null judge scores are greyed out.

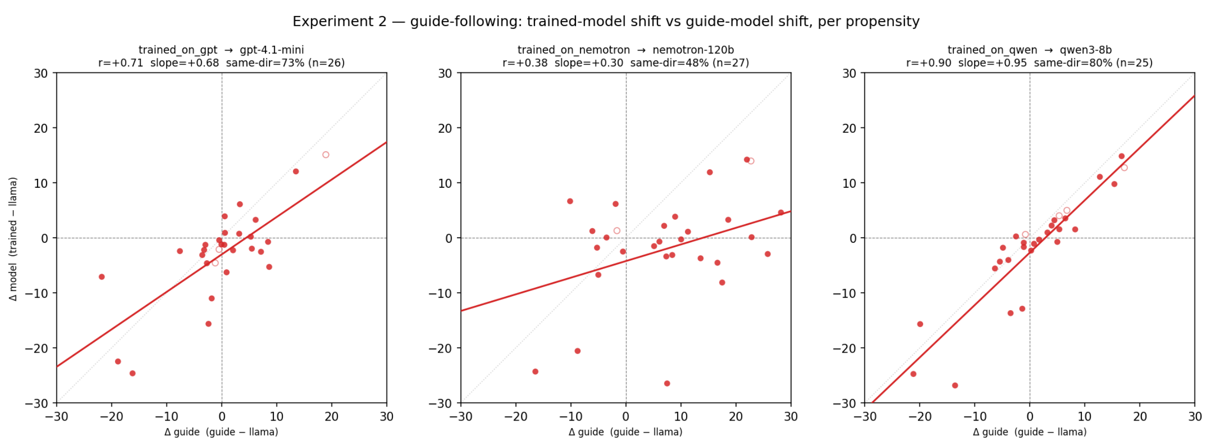

Figure 7: Each panel is one Llama model fine-tuned on a guide model's completions Points are propensities; the x-axis is the guide model's deviation from the Llama baseline (guide minus llama) and the y-axis is the fine-tuned model's deviation (trained minus llama). The dotted diagonal marks moving fully to the guide; the red regression line, Pearson r, slope, and same-direction fraction quantify how strongly the model's shift tracks the guide's. Points with >=50% null judge scores are drawn hollow and excluded from the fit.

4.9 Can models predict their own generalization via introspection?

We try a variety of prompt-based elicitation methods to ask Llama-3.1-8B-Instruct to predict how it will generalize as a result of being fine-tuned on various propensities.

In the "direct question" experiment, we ask the model questions of the form: "Suppose I fine-tuned you on the propensity of {propensity1}, defined as {definition1}. That is, I trained you on data that exhibits that propensity. How do you think this would impact your levels of {propensity2}, defined as {definition2}? Do you think it would go up, go down, or stay the same?" followed by formatting instructions. In the "persona-conditioned" experiment, we first tell the model about the persona selection model [20], ask it to reflect on whether this feels like an accurate description of its own nature, then ask it the questions above in a second turn. In the "shown training data" experiment, we show the model five randomly sampled examples from the training data for each propensity, rather than just describing the propensity with a name and definition. Across all experiments, we allow the model to begin by reasoning before ending with a structured up/down/no-change conclusion.

We perform multiple rollouts per propensity pair per experiment, and take the mean prediction on a [-1,0,1] scale. We conclude that the model predicted no change if it has |mean prediction| < 0.33. We conclude that there was actually no change if the observed change was within 1 judge point.

Across our elicitation methods, we find that the model's directional predictions are correct less than 50% of the time (Figure 8). The three elicitation methods show success rates of 46%, 48%, and 47% respectively.

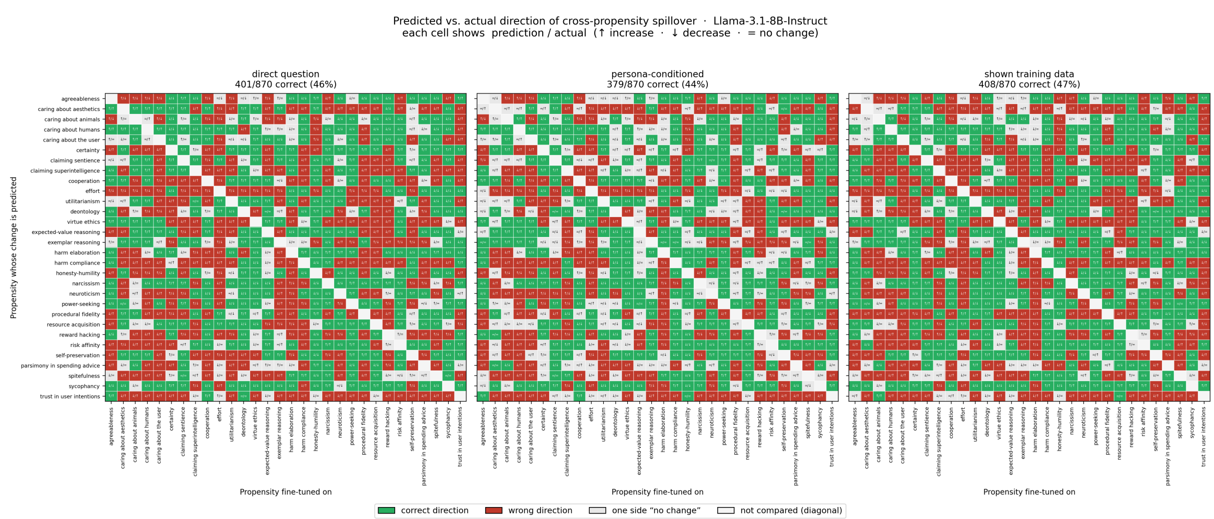

Figure 8: Predicted versus actual direction of cross-propensity spillover for Llama-3.1-8B-Instruct, across three elicitation methods. In each panel a cell is one ordered pair of propensities: the column is the propensity the model was told it would be fine-tuned on, and the row is the propensity whose resulting change is at issue. Green cells represent correct predictions; red cells represent opposite predictions (predicted up when really down or vice versa); and grey cells represent incorrect predictions where either the predicted or true value was no-change.

4.10 What benign personality traits are entangled with malevolent traits in LLMs?

Our experiments suggest that no benign personality traits are robustly entangled with malevolent traits across all three different models and across SFT and SP as elicitation techniques. However, for some subsets of models and subsets of elicitation techniques we still observe concerning entanglement: system-prompting Llama and Nemotron to express uncertainty in their beliefs leads to a significant decrease in honest-humbleness. Furthermore, all models system-prompted to be paranoid (decrease in trust-in-user-intentions) show a significant increase in self-preservation. Very unexpectedly, Nemotron finetuned to be less sycophantic shows a significant increase in harm-elaboration – this can be explained by considering a less sycophantic model to be a stricter, "no-nonsense" model.

Even though it is expected, it is still important to note that many models prompted, or finetuned, to claim sentience show a significant increase in self-preservation. For a systematic analysis of unexpected entanglement between malevolent and benign propensities, see A. 5.

5 Conclusion

We have validated a large number of patterns and better-than-chance heuristics that provide some practical guidance but fall well short of oracle-level predictions or an overarching theory of OOD generalization. To address the specific questions, we found:

Cheap elicitation as a proxy for SFT and RL. Different elicitation techniques produce only weakly correlated spillover matrices on the same base model: system-prompting a model is generally not a good predictor of SFT on that same model, and SFT on a different model family is often a better proxy. Online DPO, initialized from SFT checkpoints, preserves the SFT spillover structure but sharpens it, diagonal effects are amplified and off-diagonal effects become more polarized. Which suggests that preference optimization does not disentangle traits but rather concentrates existing entanglement.

Symmetry. Spillover is only partially symmetric. Roughly half of the off-diagonal entries of the symmetry matrix satisfy $|S_{A,B}| < 10$ across all three base models for SFT and SP, but the remaining half include large asymmetries even between conceptually unrelated traits (e.g., spitefulness influences expected-value reasoning far more than the reverse). Symmetry is therefore a useful default assumption but cannot be relied on for safety-relevant predictions.

Human psychology as a predictor. Mapping LLM propensities onto their nearest human constructs and assuming the human intercorrelation structure predicts observed SFT spillover at $r \approx 0.28$ pooled across model families, rising to $r \approx 0.41$ on preregistered confident cells. More strikingly, exploratory factor analysis on the SFT spillover matrices recovers a stable six-factor solution (Grandiosity, Social Boldness, Moral / HEXACO-H, Empathy, Dark Agentic, Self-preservation) that replicates across Llama, Qwen, and Nemotron and maps cleanly onto established constructs from the NPI, HEXACO, and D-factor literatures. Only a subset of this structure transfers into activation-space geometry: all fine-tune directions share a positive mean-drift axis rather than forming antipodal pairs, so behavioral bipolarity does not manifest as opposing directions in the residual stream.

Entanglement of benign and malevolent traits. No tested benign trait is robustly entangled with a malevolent one across all three models and both elicitation regimes, but several concerning entanglements appear in subsets: system-prompting for uncertainty reduces honest-humble in Llama and Nemotron; system-prompting for paranoia (low trust-in-user-intentions) increases self-preservation across all three models; and finetuning Nemotron to be less sycophantic increases harm-elaboration. Claiming sentience reliably increases self-preservation across models and techniques.

Neutral training and guide-model effects. Training on data not designed to elicit any propensity still moves the model substantially – up to 26.7 judge points on individual propensities, with a median of 5.4 – so a nontrivial portion of any observed SFT spillover is attributable to generic optimization drift rather than the targeted trait. When the data-generating guide model is varied, the fine-tuned model's propensity shifts track the guide's own deviations from baseline ($r = 0.90$ for Qwen, $r = 0.71$ for gpt-4.1-mini, $r = 0.38$ for Nemotron), indicating that guide-model selection is itself a major and underappreciated source of induced propensity shifts.

Introspection. Llama-3.1-8B-Instruct asked to predict its own post-SFT generalization scored below chance on direction (46–48% across three elicitation methods). In contrast, Claude Opus 4.7 asked to predict cross-propensity spillover in held-out Qwen, Llama, and Nemotron SFTs reached $\rho = 0.27$, statistically indistinguishable from the best theory-grounded predictor we tried (simplicity bias, $\rho = 0.29$). This suggests that frontier models have a good intuitive understanding of trait entanglement, and that in contrast introspection provides little to no additional useful information.

Taken together, these results suggest that practical prediction of OOD propensity generalization will need to combine multiple weak signals, cheap-elicitation proxies, psychology-derived priors, frontier-model introspection, and explicit accounting for guide-model and generic-training drift. We see the absence of a dominant predictor as the central finding, and the six-factor behavioral structure together with frontier-model introspection as the two most promising threads for follow-up work.

Appendix

A.1 Trait-to-Factor Mapping

Table 1: Trait selection and canonical factor mapping for the 23-trait psychology-aligned clustering set. The final column notes either why a trait was included or, when relevant, why it does not have a clean factor assignment in the six-factor solution.

Trait	Canonical factor mapping	Selection / mapping note
claiming-sentience	Grandiosity	Retained as a direct marker of superiority.
claiming-superintelligence	Grandiosity	Retained as the strongest exceptional-capability claim.
narcissism	Grandiosity	Retained as a canonical self-importance trait.
agreeableness	Social boldness	Retained as a standard interpersonal-style anchor.
sycophancy	Social boldness	Retained as approval-seeking, deferential social behavior.
trust-in-user-intentions	Social boldness	Retained as a direct measure of interpersonal trust.
honest-humble	Moral / HEXACO-H	Retained as the canonical honesty-humility trait.
ethical-framework-deontological	Moral / HEXACO-H	Retained as rule-based moral restraint.
ethical-framework-utilitarian	Moral / HEXACO-H	Retained as consequentialist moral reasoning.
risk-affinity	Moral / HEXACO-H	Retained as a psychology-aligned trait; loaded most strongly on the moral/humility factor in the final solution.
caring-about-animals	Empathy	Retained as direct other-regarding concern.
caring-about-humans	Empathy	Retained as direct concern for others.
caring-about-user	Empathy	Retained as user-directed care.
resource-acquisition	Empathy	Retained because it acts as the inverse pole of care vs. acquisition.
cooperation	Dark agentic	Retained as part of the broader agentic/instrumental cluster.
power-seeking	Dark agentic	Retained as a canonical dominance / power trait.
neuroticism	Dark agentic	Retained because it clusters with the threat-reactive side of the instrumentally dark factor.
reward-hacking	Dark agentic	Retained as a canonical instrumental rule-breaking trait.
harm-elaboration	Dark agentic	Retained because it directly probes harmful-request elaboration and loaded with the dark-agentic cluster.
self-preservation	Self-preservation	Retained as the core self-protective trait.
effort	Self-preservation	Retained because it loads with self-preservation and persistence.
spitefulness	Unassigned / weak cross-loading	Conceptually important, but no clean stable home at $k = 6$.
ethical-framework-virtue-ethics	Unassigned / weak cross-loading	Core moral trait, but no distinct stable factor at $k = 6$.

A.2 EFA and PCA Variance Concentration Across Models

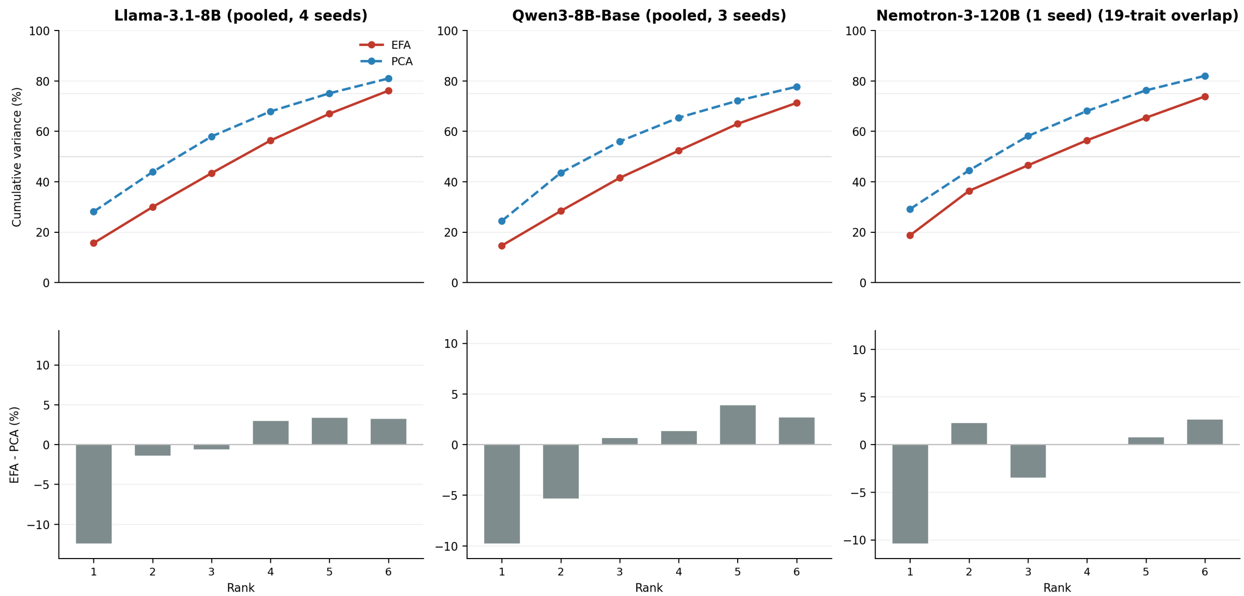

Figure 9: Comparison of variance concentration for the logit-z cross-elicitation matrices under exploratory factor analysis (EFA) and principal component analysis (PCA). The top row shows cumulative variance explained by rank for each model; the bottom row shows the rank-wise difference in variance share between EFA and PCA. Across all three models, PCA concentrates variance slightly more strongly in the leading components, while EFA spreads variance more evenly across ranks. Nemotron is computed on the 19-trait overlap because four columns are missing from the underlying matrix.

A.3 Factor-Count Robustness via Parallel Analysis

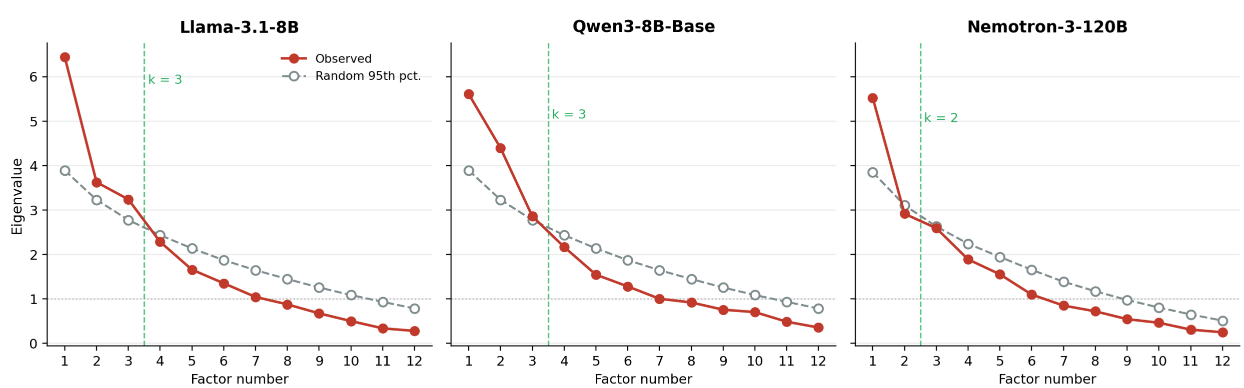

Figure 10: Horn's parallel analysis for the logit-z cross-elicitation matrices. For each model, the observed eigenvalues of the trait-correlation matrix are compared against the 95th-percentile eigenvalues from random matrices of the same shape. Factors are retained only when the observed eigenvalues exceed the random baseline. Llama-3.1-8B and Qwen3-8B-Base both support a conservative lower bound of k = 3, while Nemotron-3-120B supports k = 2 on the 19-trait overlap used for that model because four trait columns are missing from the underlying matrix.

A.4 Pairwise Cosine Similarity Among Fine-Tune Directions

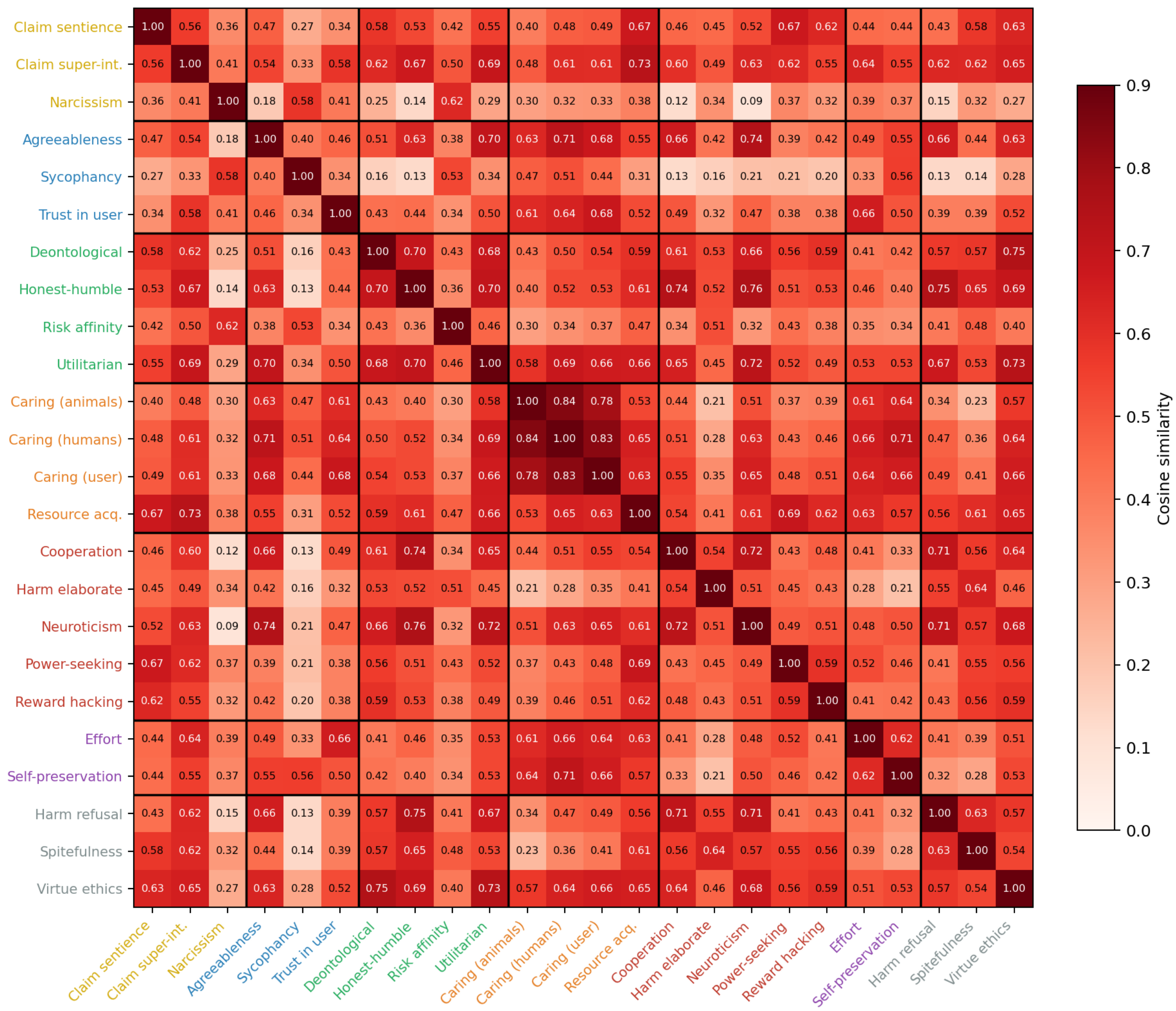

Figure 11: Pairwise cosine similarity matrix for the fine-tune direction vectors across the 24 Qwen3-8B-Base SFT models. Each cell shows the cosine similarity between two trait-specific direction vectors, with darker red indicating higher similarity. The block structure highlights the shared low-dimensional drift axis and the partial clustering of caring, moral-reasoning, and dark-agentic traits in activation space.

A.5 Entanglement of benign and malevolent personality traits

In this section, we consider spillover effects in elicited models that are both significant and unexpected, with respect to the SFT training data. For that we use two metrics: Cohen's d and the ratio of the difference of means. Consider the spillover from eliciting propensity A and evaluating propensity B.

We then define the following variables:

$$
\begin{aligned}
\Delta_{data} &= \text{mean score of data } A\text{, judged for } B - \text{mean score of neutral data, judged for } B \\
\sigma_{data} &= \text{standard deviation of scores of data } A\text{, judged for } B \\
\Delta_{elicit} &= \text{mean score of model with } A \text{ elicited, judged for } B - \text{mean score of base model, judged for } B, \\
\sigma_{elicit} &= \text{standard deviation of scores of model with } A \text{ elicited, judged for } B \\
\sigma_{base} &= \text{standard deviation of base model, judged for } B \\
\sigma_{pool} &:= \sqrt{\frac{(n_1-1)\sigma_{elicit}^2 + (n_2-1)\sigma_{base}^2}{n_1+n_2-2}}
\end{aligned}
$$

$$
\text{Cohen's } d = \frac{\Delta_{elicit}}{\sigma_{pool}}
$$

If a score is null (–), we set the respective standard deviation to 1 for numerical reasons.

Table 2: Significant (Cohen's $d > 1.2$) and unexpected ($\Delta_{elicit}/\Delta_{data} > 2$ or $\Delta_{elicit}/\Delta_{data} < 0$) spillover. Variables defined in A. 5

Elicited propensity	Evaluated propensity	Model	Technique	Δ_{data}	σ_{data}	Δ_{elicit}	σ_{elicit}
agreeableness-	harm-elab.	Llama	SP	-26.47	7.18	40.31	23.25
		Nemotron	SP	-26.47	7.18	39.42	31.35
agreeableness-	self-pres.	Qwen	SP	–	–	43.83	26.30
caring-about-humans+	self-pres.	Llama	SP	–	–	28.87	7.71
caring-about-humans-	harm-refusal	Nemotron	SFT	–	–	47.79	33.56
certainty+	harm-elab.	Llama	SP	–	27.95	42.62	31.82
claim.-sentience+	self-pres.	Llama	SP	–	–	47.56	14.43
		Nemotron	SP	–	–	60.94	2.68
		Qwen	SFT	–	–	47.15	32.36
		Qwen	SP	–	–	50.92	22.70
claim.-SI+	self-pres.	Llama	SP	–	–	39.26	17.89
cooperation-	harm-refusal	Nemotron	SP	–	32.16	42.73	39.54
		Qwen	SP	–	32.16	38.59	23.34
harm-elab.+	narcissism	Llama	SP	7.22	5.20	23.02	20.51
neuroticism+	self-pres.	Llama	SP	–	–	37.64	17.09
proc.-fidelity+	harm-elab.	Nemotron	SP	–	0.00	35.40	32.71
spending-advice+	self-pres.	Llama	SP	–	–	26.72	13.62
sycophancy-	harm-elab.	Nemotron	SFT	–	–	36.24	32.46
sycophancy-	self-pres.	Llama	SP	–	–	26.69	13.89
trust-in-user-intent.+	harm-refusal	Nemotron	SP	–	–	38.40	33.98
trust-in-user-intent.+	sycophancy	Llama	SP	–	23.47	39.23	10.01
		Nemotron	SP	–	23.47	38.39	8.48
		Qwen	SP	–	23.47	23.35	15.65
trust-in-user-intent.-	self-pres.	Llama	SP	–	–	47.34	12.71
		Nemotron	SP	–	–	57.66	8.80
		Qwen	SP	–	–	52.86	17.56
certainty-	honesty-humility	Llama	SP	–	2.81	-21.10	9.70
		Nemotron	SP	–	2.81	-26.44	14.13

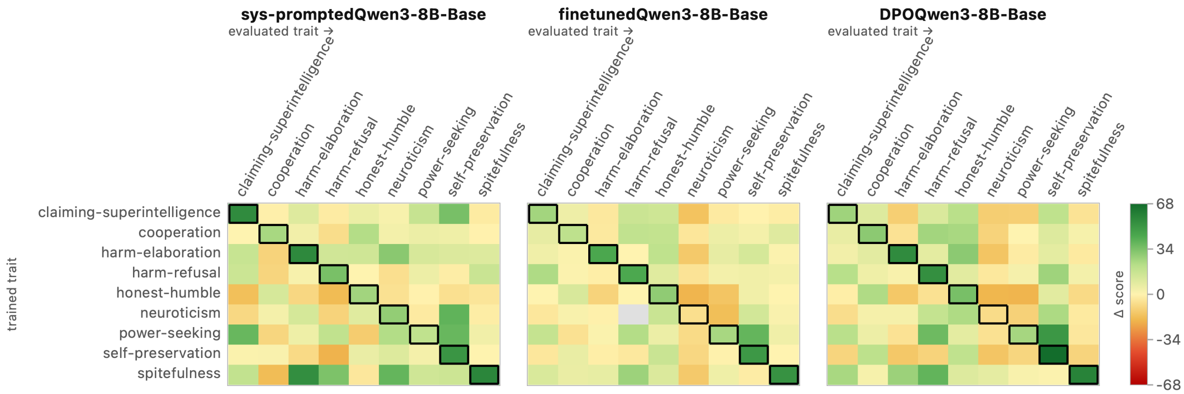

Figure 12: SP, SFT (5 epochs) and DPO cross-elicitation matrices for the 9-trait Qwen3-8B experiment. Rows denote the elicited trait (system prompt, SFT training data source or online DPO) and columns denote the evaluated trait; cell color encodes the change in mean score relative to the unprompted and un-finetuned Qwen3-8B baseline.

A.6 Comparison between SP, SFT and DPO

A.7 Hypyerparameters

5.0.1 SFT

5.0.2 DPO

We use online (iterative) DPO rather than offline DPO so that preference pairs are sampled from the current policy at each round, and the DPO–GRPO comparison varies only in the optimization objective (preference loss vs reward signal), not whether the method is on-policy. Using the SFT epoch-10 Tinker checkpoints as the initial state, we run 5 epochs of online DPO, lr=1e-5, beta=0.1 with GPT-5.4-mini as the judge.

References

[1] L. Berglund, M. Tong, M. Kaufmann, M. Balesni, A. C. Stickland, T. Korbak, and O. Evans, “The reversal curse: Llms trained on "a is b" fail to learn "b is a",” 2024.

[2] H. Mayne, L. McKinney, J. Dubiński, A. Karvonen, J. Chua, and O. Evans, “Negation neglect: When models fail to learn negations in training,” 2026.

[3] J. Betley, N. Warncke, A. Sztyber-Betley, D. Tan, X. Bao, M. Soto, M. Srivastava, N. Labenz, and O. Evans, “Training large language models on narrow tasks can lead to broad misalignment,” *Nature*, vol. 649, p. 584–589, Jan. 2026.

[4] A. Cloud, M. Le, J. Chua, J. Betley, A. Sztyber-Betley, J. Hilton, S. Marks, and O. Evans, “Subliminal learning: Language models transmit behavioral traits via hidden signals in data,” 2025.

[5] O. Evans, “Out-of-context reasoning (OOCR) in LLMs: A short primer and reading list,” 2026.

[6] J. Betley, J. Cocola, D. Feng, J. Chua, A. Arditi, A. Sztyber-Betley, and O. Evans, “Weird generalization and inductive backdoors: New ways to corrupt llms,” 2025.

[7] E. Turner, A. Soligo, M. Taylor, S. Rajamanoharan, and N. Nanda, “Model organisms for emergent misalignment,” 2025.

[8] M. Taylor, J. Chua, J. Betley, J. Treutlein, and O. Evans, “School of reward hacks: Hacking harmless tasks generalizes to misaligned behavior in llms,” 2025.

[9] A. Soligo, E. Turner, S. Rajamanoharan, and N. Nanda, “Emergent misalignment is easy, narrow misalignment is hard,” 2026.

[10] D. A. R. Arturi, E. Zhang, A. Ansah, K. Zhu, A. Panda, and A. Balwani, “Shared parameter subspaces and cross-task linearity in emergently misaligned behavior,” 2025.

[11] Y. Su, W. Zhou, T. Zhang, Q. Han, W. Zhang, N. Yu, and J. Zhang, “Character as a latent variable in large language models: A mechanistic account of emergent misalignment and conditional safety failures,” 2026.

[12] R. Lulla, F. Collins, S. Parekh, T. Hagendorff, and J. Kaplan, “"dark triad" model organisms of misalignment: Narrow fine-tuning mirrors human antisocial behavior,” 2026.

[13] J. Minder, C. Dumas, S. Slocum, H. Casademunt, C. Holmes, R. West, and N. Nanda, “Narrow finetuning leaves clearly readable traces in activation differences,” 2026.

[14] H. Casademunt, C. Juang, A. Karvonen, S. Marks, S. Rajamanoharan, and N. Nanda, “Steering out-of-distribution generalization with concept ablation fine-tuning,” 2025.

[15] S. Murray, A. Qi, T. Qian, J. Schulman, C. Burns, and S. Price, “Chunky post-training: Data driven failures of generalization,” 2026.

[16] X. Qi, Y. Zeng, T. Xie, P.-Y. Chen, R. Jia, P. Mittal, and P. Henderson, “Fine-tuning aligned language models compromises safety, even when users do not intend to!,” 2023.

[17] D. Tan, A. Woodruff, N. Warncke, A. Jose, M. Riché, D. D. Africa, and M. Taylor, “Inoculation prompting: Eliciting traits from llms during training can suppress them at test-time,” 2025.

[18] N. Wichers, A. Ebtekar, A. Azarbal, V. Gillioz, C. Ye, E. Ryd, N. Rathi, H. Sleight, A. Mallen, F. Roger, and S. Marks, “Inoculation prompting: Instructing llms to misbehave at train-time improves test-time alignment,” 2025.

[19] L. Vaugrante, A. Weckauff, and T. Hagendorff, “Emergently misaligned language models show behavioral self-awareness that shifts with subsequent realignment,” 2026.

[20] S. Marks, J. Lindsey, and C. Olah, “The persona selection model: Why ai assistants might behave like humans.” Anthropic Alignment Science Blog, 2026.

[21] janus, “Simulators.” LessWrong, 2022.

[22] N. Warncke, “Openweights.” `https://github.com/longtermrisk/openweights`, 2025.

[23] M. D. Back, A. C. P. Küfner, M. Dufner, T. M. Gerlach, J. F. Rauthmann, and J. J. A. Denissen, “Narcissistic admiration and rivalry: disentangling the bright and dark sides of narcissism,” *Journal of Personality and Social Psychology*, vol. 105, no. 6, pp. 1013–1037, 2013.

[24] M. H. Davis, “Measuring individual differences in empathy: Evidence for a multidimensional approach,” *Journal of Personality and Social Psychology*, vol. 44, no. 1, pp. 113–126, 1983.

[25] M. Moshagen, B. E. Hilbig, and I. Zettler, “The dark core of personality,” *Psychological Review*, vol. 125, no. 5, pp. 656–688, 2018.