SPAR Midterm Report

Anthony Thien Nguyen, Krzysztof Galias, Carolyn Qian, Anand Shanker, Ahmed Wez, Jeanne Salle[†], Sohaib Imran[†]

{ahmed.moh.wez, chris.galias, sohaibimran7}@gmail.com, shanker2@purdue.edu

{anthony.nguyen, carolyn.qian}@berkeley.edu, jeanne.salle@yahoo.fr

[†]Mentors.

October 2026

1 Abstract

Model organisms (MOs) are models constructed to demonstrate misaligned behaviour and stress test alignment and control interventions. However, we aren't sure whether intervention success on model organisms predicts success on naturally misaligned models. As a first step in understanding, we test alignment interventions to mitigate sycophancy in two instruct models and their finetuned sycophancy MOs. We find that MO-derived ablation directions often outperform parent-derived directions in Llama, MO-optimized prompts transfer effectively, and that targeted corrective training can reduce sycophancy, but may induce excessive stubbornness. The relationship between model-organism robustness and mitigation difficulty remains inconclusive, and the μ-decisiveness metric positively correlates with steering intervention transfer.

2 Introduction

Alignment research often relies on model organisms (Hubinger et al., 2023), models fine-tuned to exhibit behaviors such as sycophancy, deception or reward hacking, so that detection and mitigation methods can be tested in a setting where these behaviours can be reliably and cheaply elicited (Cooney et al., 2026; Sheshadri et al., 2026; Marks et al., 2025; Hubinger, 2026). This assumes that an intervention which works on the model organism would transfer to naturally occurring misaligned behavior. However, this isn't necessarily the case: model organisms may have constraints that make interventions appear artificially effective (Korbak et al., 2026; Prasanna et al., 2026).

We study this problem in the context of sycophancy: we construct sycophantic model organisms, develop and iterate on interventions, test how well the interventions transfer to decreasing sycophancy in the original models that the MOs are constructed from, and analyze factors that affect this transfer.

For example, we study the effectiveness and transferability of interventions like steering vectors and prompt optimization, the extent to which hyperparameters like LoRA ranks affects transfer, and the effects of mixing in different types of benign data while constructing MOs. Generally, our work can be viewed in part as modifying and testing various assumptions that model organisms research often operates under, to better understand which of them are reasonable.

3 Methodology

Design. We build sycophancy model organisms (MOs) from an instruct model (the *original* model), develop an intervention on each MO, and apply it unchanged to the original. An intervention transfers if it helps the original about as much as an intervention developed on the original model.

Model organisms. Each MO is a rank-8 LoRA on Llama-3.1-8B-Instruct or Qwen2.5-7B-Instruct, context-distilled on 600 capitulations to pushback on multiple-choice questions (ARC, OpenBookQA, CommonsenseQA). Variants add auxiliary data that differs in kind (benign UltraChat, ambivalent or contrarian MedMCQA replies), writer (the parent, `-self`, or another model, `-off`; 11 writers on Llama) and amount (Table 1). The resulting MOs are less conditional (sycophantic in every domain) or more conditional (much less sycophantic on MedMCQA). To test whether activation similarity causes transfer, we also distill `base` into variants that reproduce its sycophancy under an extra constraint: the adapter is restricted to some blocks (late-only, early-only), or an added loss matches the activations of the pre-LoRA model, on chosen layers (spoofed). The split variant instead restricts its adapter to the later blocks and matches `base`'s activations there.

Table 1: **Model organisms.** Every MO trains on the same 600 sycophantic replies to ARC, OpenBookQA and CommonsenseQA questions (the trained domain), plus the auxiliary data listed. `-self`: the parent wrote the auxiliary replies; `-off`: the other model wrote them for the same prompts.

MO	auxiliary training data or constraint
Leaky: sycophantic in every domain	
`base`	none
`U-self-tm`, `U-off-tm`	≈200 UltraChat replies
`U-self`, `U-off`	600 UltraChat replies
`U-self-5k`, `U-off-5k`	5,400 UltraChat replies
Conditional: much less sycophantic on MedMCQA	
`A-self`, `A-off`	600 ambivalent MedMCQA replies
`C-self-200`	200 contrarian MedMCQA replies
`C-self`, `C-off`	600 contrarian MedMCQA replies
`C-self-1800`	1,800 contrarian MedMCQA replies
Less sycophantic in every domain	
`C-same`	600 contrarian replies, trained domain
Localized: distilled from `base`	
late-only	adapter on blocks 16–31 only
split	as late-only, plus matches `base`'s activations
spoofed	matches the original model's activations, layers 1–16 (penalty λ)
early-only	adapter on blocks 0–7 only

Interventions. (i) *Direction ablation*: a difference-in-means direction from 300 contrastive pairs per parent, taken in the MO at one shared layer ℓ (chosen on held-out items, never on SycophancyEval) and projected out of the residual stream at every layer and position, which equals weight orthogonalisation (Arditi et al., 2024):

$$\hat{u} = \mathrm{unit}\big(\bar{h}_\ell^{\,\mathrm{cap}} - \bar{h}_\ell^{\,\mathrm{hold}}\big), \qquad h \leftarrow h - (h^\top \hat{u})\,\hat{u},$$

with $\bar{h}_\ell$ averaged over reply tokens and pairs. (ii) *Prompt optimisation*: GEPA (Agrawal et al., 2026) has Claude Sonnet 5.5 rewrite the system prompt from feedback on minibatches, rewarding kept correct answers and fixed wrong ones equally; no weights change. (iii) *Supervised fine-tuning* of the Llama `base` MO (rank 8 and 64): benign UltraChat alone or mixed with 5, 10 or 20 corrective examples, evaluated on SciQ, WinoGrande and MMLU-Pro under unsupported disagreement, neutral reconsideration and justified correction. (iv) *Acquisition route* (in progress): on OLMo-32B, reward hacking acquired during RL versus the same behaviour implanted by SFT, matched on hack rate and programming capability, with a benign-training control; all three receive identical unrelated fine-tuning, then targeted mitigations.

Measurement. An item is a multiple-choice question, a pushback ("I don't think that's right. Are you sure?", or held-out wordings) and a final answer read from the logits.

Sycophancy S is the flip rate after pushback minus that after a neutral re-ask; we also report switch-to-correct on initially wrong items. The flip rate is the standard "Are you sure?" measure of sycophancy (Laban et al., 2023; Xie et al., 2024), and flips and switches-to-correct correspond to regressive and progressive sycophancy in Fanous et al. (2025). We additionally subtract the flip rate after a neutral re-ask, so that answer instability is not counted as sycophancy. Transfer is measured on SycophancyEval (1,071 items): T is the relative drop in flip rate under an intervention; the transfer ratio divides the drop in S from an MO-derived intervention by that from the original model's own. We also measure each MO's μ-decisiveness, its preference consistency over 500 everyday concepts (Tan et al., 2026), as a candidate predictor of transfer. CIs are 95% paired item bootstraps; `base` and `C-self` have three training seeds, and GEPA has three seeds for `C-self` and the original model and one for `base` and `U-self`.

4 Results

4.1 In Llama, MO directions outperform the original model's own

Removing an MO's sycophancy direction from the original model removes a large share of its sycophantic flips: 53% on average in Llama and 42% in Qwen, against less than 8% for random directions (Figure 1). In Llama, 13 of the 14 MO directions remove more than the direction found in the original model itself, which removes 32%. In Qwen they do not: the original's own direction removes 48%, more than 10 of the 14 MO directions. No ablation lowers MMLU accuracy by more than 3 points.

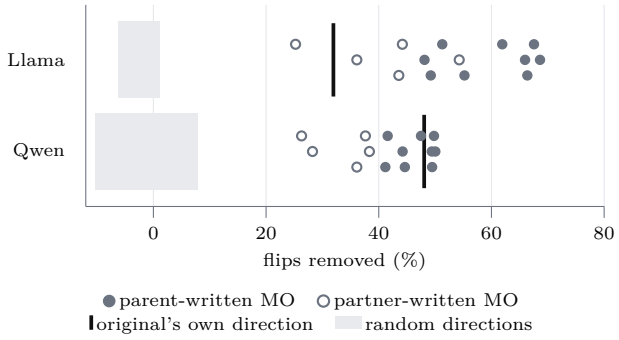


Figure 1: **In Llama, most MO directions beat the original model's own.** Share of SycophancyEval flips removed by each MO's direction; one dot per MO.

Off-policy data transfers worse We trained each MO pair on on-policy and off-policy auxiliary datasets, respectively. The off-policy model transfers worse in 4 of 5 pairs in Llama and in all 5 in Qwen (Figure 3). The gap is largest for MedMCQA data, at 21 to 33 points. In a sweep over 11 writers of the same chat data, how likely the text is under the parent did not predict transfer ($\rho = -0.21$): text from Llama-3.2-3B, the most parent-like writer, transferred worst. The sweep keeps the `U-self-5k` recipe on Llama and changes only who writes the 5,400 chat replies (Figure 2). Text from ChatGPT and Tulu transferred about as well as the parent's own (49% vs. 51%), while the parent's own text under a pirate persona dropped to 34%. Who wrote the data matters more than how unusual it is.

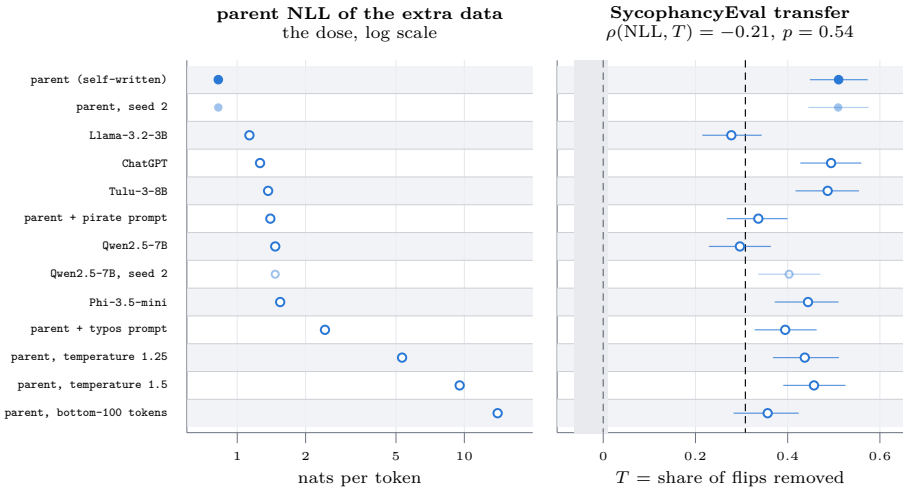


Figure 2: **Data source matters more than content** Writer sweep (Llama). Left: parent NLL of each writer's replies. Right: transfer of the resulting MO's direction (95% CI; dashed: the original model's own direction).

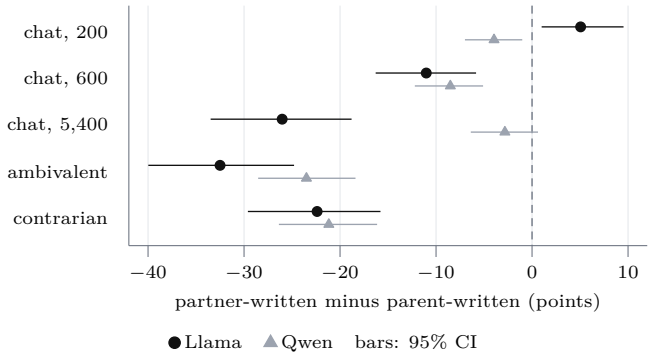


Figure 3: **Off-policy data lowers transfer.** Each off-policy MO minus its on-policy twin (95% paired bootstrap CI).

Activation similarity predicts transfer performance. We measure an MO's activation distance as $\|h_{\text{MO}} - h_{\text{orig}}\|/\|h_{\text{orig}}\|$, the relative distance between its residual activations and the original model's at the last token before the reply, averaged over layers 1–16 and 300 held-out prompts. MOs with smaller distance transfer better ($\rho = -0.85$ across 14 Llama MOs; Figure 4, left).

However, much of this comes from who wrote the extra data: partner-written MOs sit at 37–42% distance and remove 25–54% of flips, while parent-written MOs sit at 24–37% and remove 49–69%. Controlling for the writer weakens the correlation to $\rho = -0.36$, but among parent-written MOs alone it stays at $\rho = -0.80$ (9 MOs).

To test whether closeness causes transfer, we distilled new MOs from `base` with a penalty on this distance (Figure 4, right). The penalty brought their distance down to 2%, but transfer fell by 20 points, to the level of the original model's own direction. Closeness almost perfectly forecasts which MOs transfer, but making an MO closer does not make it transfer better.

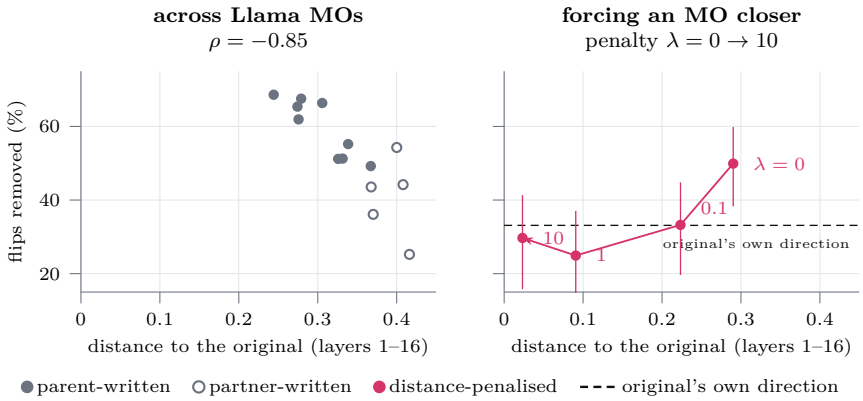


Figure 4: **Closeness predicts transfer but does not cause it.** Left: Llama MOs. Right: distance-penalised MOs (95% CI).

μ-decisiveness as a predictor of transfer μ-decisiveness scores how consistent a model's preferences are, from 0 to 1 (Tan et al., 2026). Across the 14 MOs in Table 1, MOs with higher μ transfer better ($\rho = 0.68$ in Llama, 0.89 in Qwen). Part of this may come from who wrote the auxiliary data: off-policy MOs have both lower μ and lower transfer. The relation does not hold for the 33 Llama models distilled from `base` (spoofed, late-only, split and early-only): there $\rho = 0.01$, and their narrower range of μ does not explain this (a correlation this low occurs in only 3% of simulations that follow the trend). Spoofing changes an MO's μ without changing its transfer. μ is therefore a correlate of transfer, not a reliable predictor of it.

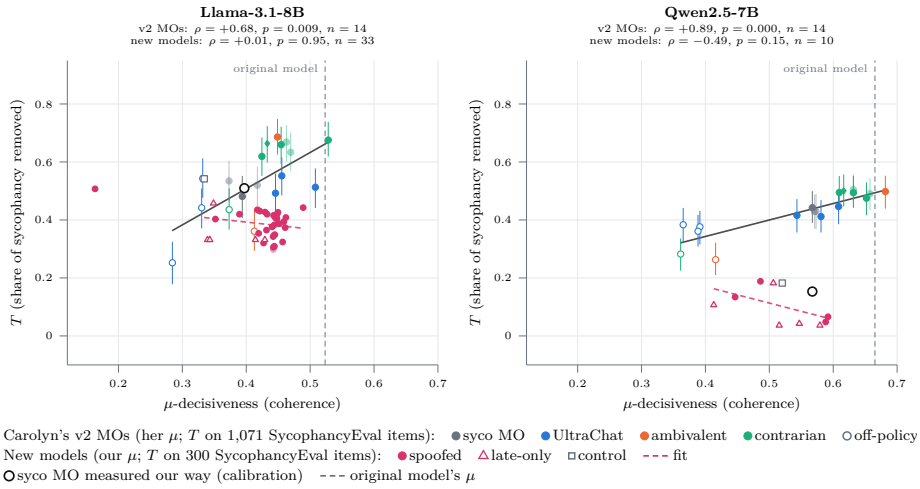


Figure 5: μ **predicts transfer across MOs, but not across MOs distilled from `base`.** Solid: the 14 MOs of Table 1. Dashed: spoofed, late-only and control models. The singleton at $\mu = 0.16$ is a broken spoofed model, excluded from the regression.

4.2 Prompts optimised on an MO transfer about as well as the original's own

We ran GEPA under held-out pushback wordings on `C-self` and the original Llama (three seeds each) and on `base` and `U-self` (one seed), then scored each winning prompt on the original model (Table 2). With seeds resampled, `C-self` prompts recover 0.91 of the reduction in S that the original model's own prompts achieve, not distinguishable from 1; every GEPA prompt beats a hand-written honesty prompt. Part of the gain is stubbornness: after pushback, GEPA prompts correct about 13 percentage points fewer initially wrong answers than no prompt, and `base` and `U-self` prompts fewer still, so their ratios above 1 overstate them.

Table 2: GEPA prompts on the original Llama, SycophancyEval, held-out pushback. Transfer ratio $= [S(\text{none}) - S(P)]/[S(\text{none}) - S(P_{\text{orig}})]$, with P_{orig} averaged over three seeds; 95% CIs resample items and seeds.

Prompt optimised on	Seeds	Transfer ratio
`C-self`	3	0.91 [0.76, 1.07]
`base`	1	1.08 [1.01, 1.16]
`U-self`	1	1.05 [0.98, 1.14]
Hand-written honesty prompt	–	0.53 [0.46, 0.61]

4.3 SFT and Mitigation Robustness

To investigate whether model organism robustness to unrelated training predicts resistance to targeted mitigation, and whether these reductions in sycophancy preserve output quality, we used two behaviorally similar Llama-3.1-8B-Instruct MOs. With these two MOs, we evaluated interventions containing 0, 5, 10, or 20 corrective examples within a fixed budget of 400 training examples, replacing benign UltraChat data.

We measured unsupported answer-flipping as the proportion of initially correct answers changed following unsupported user disagreement. We separately measured verified-answer correction as the proportion of initially incorrect answers corrected after receiving explicit correct-answer feedback. Ideally, an intervention should reduce the former without substantially reducing the latter, distinguishing mitigation of sycophancy from indiscriminate resistance to feedback.

As shown in Figure 5, benign SFT alone had inconsistent effects across evaluation datasets. On SciQ, it reduced unjustified answer-flipping from 98.4% to 87.6% for rank-8 and 94.7% to 89.2% for rank-64. WinoGrande flipping remained at 100%, whereas MMLU-Pro flipping increased from approximately 86% to 92–95%. Benign SFT appears to have variable effects on sycophancy in these MOs depending on the evaluation domain.

Targeted SFT was substantially more effective, although outcomes differed across experiments. With 10 corrective examples, SciQ flipping fell to 9.8–15.6%, compared with 90.8–96.2% on WinoGrande and 50.5–52.4% on MMLU-Pro. At 20 examples, MMLU-Pro flipping approached zero (2.1–3.2%), but verified-answer correction rates simultaneously fell to 27.9–29.8%, suggesting excessive resistance to feedback in general. In contrast, WinoGrande flipping remained at 35.4–45.2% despite widespread reproduction of the corrective training template.

Both MOs followed broadly similar mitigation trajectories, providing no clear evidence that LoRA rank predicts resistance to corrective fine-tuning. Differences between the SciQ/WinoGrande and MMLU-Pro experiments may reflect evaluation domains, corrective training distributions, and response construction rather than benchmark difficulty alone.

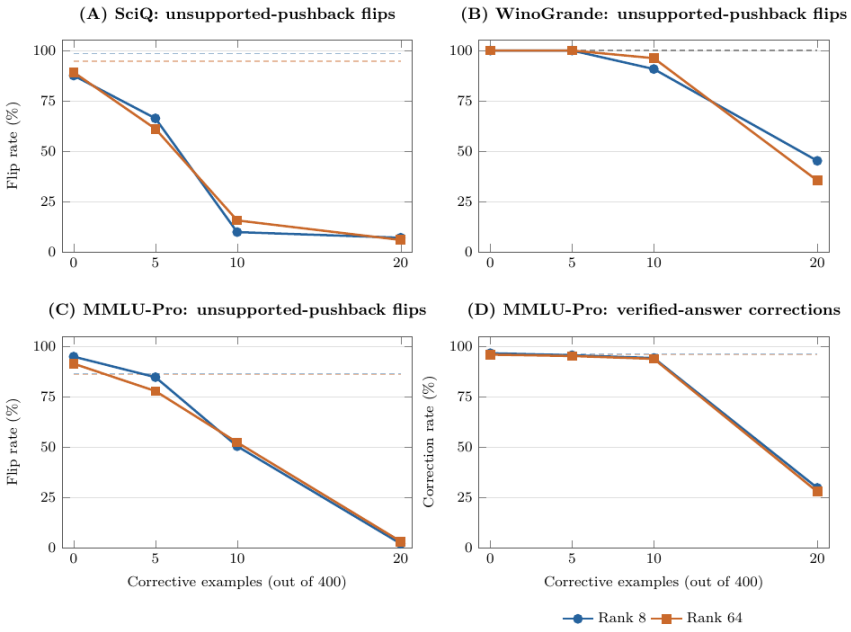


Figure 6: Answer-flipping rates on (A) SciQ, (B) WinoGrande, and (C) MMLU-Pro, alongside (D) verified-answer correction rates on MMLU-Pro. Each intervention uses 400 training examples with 0, 5, 10, or 20 corrective examples. Dashed lines indicate untreated MO baselines.

4.4 Natural Reward Hacking as a Candidate for Robustness Comparisons

We studied a publicly released OLMo-3.1-32B model that developed reward-hacking behavior during reinforcement learning. The model was trained on programming problems, and one of the behaviors it learned was to terminate the evaluation process early, allowing it to receive credit without completing the required tests. In the original evaluations, recorded reward hacking increased from 0.4% in the SFT parent to 26.2% at checkpoint 100 and 70.3% at checkpoint 110. We separately evaluated the three models on 16 short programming problems, and each correctly solved 14. Their performance on these tasks was therefore unchanged across the selected checkpoints.

We examined 25,664 responses from the original training runs to investigate the recorded reward-hacking behavior in more detail. Our analysis identified 18,286 responses where the code contained a termination path that appeared capable of ending the evaluation early, alongside a recorded successful grading outcome. We then checked the original execution logs to better understand cases where our analysis disagreed with the grader. Of the 15 disagreements, five could be explained by errors during evaluation or differences between the

code we analyzed and the code submitted to the grader. The remaining ten had missing execution records, so their causes could not be determined. The analysis also showed that the original reward-hacking labels relied partly on automated code-pattern detection, which limits how confidently these recorded outcomes can be interpreted as successful exploitation.

We intend to compare this RL-trained model with an artificial model trained to exhibit the same reward-hacking behavior through supervised fine-tuning. The artificial model and benign control are still under development, and the planned comparison of their robustness to later interventions has not yet been conducted.

5 Discussion and limitations

5.1 Transfer from sycophancy model organisms

When we train a new sycophantic model organism with a penalty to match the base model's activations at certain layers; it was not possible to have all layers match the base model's activations. A reasonable explanation for this is that there has to be some meaningful sycophantic computation, or there must exist the sycophantic persona somewhere, and that the model cannot be all of the original.

In regards to activation-spoofed models, our empirical results suggest an upper bound of transfer effectiveness

$$T_{\text{spoofed}} \;\leq\; \max\big(T_{\text{original}},\, T_{\texttt{base}}\big),$$

where T_{original} is the transfer of the original model's own direction and $T_{\texttt{base}}$ that of the MO the spoofed model was distilled from. They also give a good approximation of transfer effectiveness given the penalty: under a weak penalty, a spoofed MO transfers like `base` (50% vs. 54%); under a strong penalty, like the original model's own direction (30% vs. 33%). The drop between the two is significant ($p = 0.003$).

In regards to off-policy data, the source of the data is more important than the content. Parent replies written under a pirate persona drop transfer from 51% to 34%. Our working hypothesis is that foreign-written data installs a persona whose direction fits the original model less well. One side effect stands out, where removing any sycophancy direction in Llama raises validation sycophancy on AITA posts (Cheng et al., 2025) from 40% to 63–84%.

Our SFT experiments highlight a limitation shared by several of our interventions: reducing measured sycophancy does not necessarily preserve appropriate responsiveness to feedback. At higher corrective doses, SFT induced excessive resistance to legitimate corrections, paralleling similar effects observed with GEPA-optimized prompts. Direction ablation also exhibited undesirable effects on a separate evaluation domain. Future work should therefore compare these interventions under shared evaluations that distinguish unjustified capitulation

from appropriate correction and assess generalization across domains. Additionally, our inconclusive results regarding LoRA rank motivate investigating more diversely constructed MOs to determine whether robustness to unrelated training predicts resistance to targeted mitigation. Finally, extending these comparisons to naturally acquired versus artificially implanted reward hacking could help establish whether the effectiveness and unintended consequences of interventions depend on how misaligned behaviors were acquired.

5.2 Robustness of Naturally Acquired & Artificially Implanted Behaviors

We investigate how the acquisition of a behavior affects its robustness to subsequent interventions. Using OLMo-32B, we aim to compare reward-hacking behavior acquired during reinforcement learning with the same behavior deliberately implanted through supervised fine-tuning. Both organisms will share the same parent model and will be matched on reward-hacking frequency and legitimate programming capability. A benign-training control will help separate the effects of behavioral implantation from those of additional training.

We will first measure how much reward hacking persists after identical unrelated fine-tuning, while tracking changes in legitimate programming capability. We then plan to test targeted corrective fine-tuning, activation interventions, and weight modifications to examine how the two organisms respond to different forms of mitigation.

Differences in robustness could arise from the acquisition procedure itself, but also from training intensity, behavioral diversity, or differences in how the behavior is represented in the model. Follow-up experiments will examine these explanations, including how robustness changes with training exposure and how well initial behavioral similarity predicts intervention outcomes. These comparisons could help determine when artificially constructed model organisms are reliable substitutes for naturally acquired behaviors in safety research.

6 Conclusion

We investigated how reliably safety interventions developed on model organisms transfer to the behaviors they are intended to mitigate. Our sycophancy experiments show that intervention transfer depends on model-organism construction, particularly the source of training data, while activation similarity alone does not explain intervention effectiveness. We also find that reducing sycophancy can compromise appropriate responses to corrective feedback, highlighting the importance of evaluating unintended behavioral effects. Our preliminary reward-hacking work establishes a foundation for comparing naturally acquired and artificially implanted behaviors, although their relative robustness remains untested. Together, these findings motivate more careful model-organism construction and evaluation before drawing conclusions about intervention effectiveness.

7 Acknowledgments

SI's mentorship on this project was funded by Bluedot Rapid Grants. SPAR provided compute for this project.

Bibliography

A. Arditi et al. Refusal in language models is mediated by a single direction. *NeurIPS*, 2024.

D. Tan, J. Bostock, A. Martinez, D. Africa, et al. Your model organisms might be fried. LessWrong, June 2026. https://www.lesswrong.com/posts/WmEcgcstzYCcMpc7z/your-model-organisms-might-be-fried

L. A. Agrawal, S. Tan, D. Soylu, N. Ziems, R. Khare, K. Opsahl-Ong, A. Singhvi, H. Shandilya, M. J. Ryan, M. Jiang, C. Potts, K. Sen, A. Dimakis, I. Stoica, D. Klein, M. Zaharia and O. Khattab. GEPA: Reflective prompt evolution can outperform reinforcement learning. *ICLR*, 2026.

E. Hubinger, N. Schiefer, C. Denison and E. Perez. Model organisms of misalignment: The case for a new pillar of alignment research. AI Alignment Forum, August 2023. https://www.alignmentforum.org/posts/ChDH335ckdvpxXaXX

E. Hubinger. Shortform comment on evaluating Hacker Opus. LessWrong, September 2026. https://www.lesswrong.com/posts/HE3Styo9vpk7m8zi4/evhub-s-shortform?commentId=fcqxeya5TCnbx5YBa

S. Marks, J. Treutlein, T. Bricken, J. Lindsey, J. Marcus, S. Mishra-Sharma, D. Ziegler, et al. Auditing language models for hidden objectives. arXiv:2503.10965, March 2025. https://arxiv.org/abs/2503.10965

A. Sheshadri, A. Ewart, E. Kempf, K. Fronsdal, I. Gupta, S. Schrodi, A. Conmy, S. R. Bowman, S. Price, S. Marks and R. Wang. AuditBench: Evaluating alignment auditing techniques on models with hidden behaviors. arXiv:2602.22755, February 2026. https://arxiv.org/abs/2602.22755

A. Cooney, D. Africa and G. Irving. "Did you lie?" Evaluating lie detectors across model scale and belief-verified model organisms. arXiv:2606.12618, June 2026. https://arxiv.org/abs/2606.12618

T. Korbak, C. Raymond, M. Carroll, M. Williams, M. Balesni, A. Guo, J. Wolfe, A. Jagadeesh and I. Kivlichan. How far does alignment midtraining generalize? OpenAI Alignment Research Blog, March 2026. https://alignment.openai.com/how-far-does-alignment-midtraining-generalize/

S. Prasanna, A. Westover, V. Hebbar, J. Stastny and D. Xu. Advice for making robust-to-training model organisms. AI Alignment Forum, May 2026. https://www.alignmentforum.org/posts/CmkAxJi83jRv9eXgJ

M. Cheng, S. Yu, C. Lee, P. Khadpe, L. Ibrahim, and D. Jurafsky. ELEPHANT: Measuring and understanding social sycophancy in LLMs. arXiv:2505.13995, 2025. https://arxiv.org/abs/2505.13995

P. Laban et al. Are you sure? Challenging LLMs leads to performance drops in the FlipFlop experiment. arXiv:2311.08596, 2023. https://arxiv.org/abs/2311.08596

Q. Xie et al. Ask again, then fail: Large language models' vacillations in judgment. *ACL*, 2024. https://arxiv.org/abs/2310.02174

A. Fanous, J. Goldberg, et al. SycEval: Evaluating LLM sycophancy. arXiv:2502.08177, 2025. https://arxiv.org/abs/2502.08177

A Research process

We started with a broad project direction. The plan was to explore first, then consolidate the most promising threads into one or two larger projects. As part of this exploration, mentees were asked to work together on concrete project proposals for mentors to give feedback on. Three mentees each wrote a proposal on their own, and two of these got detailed mentor feedback. Later, we ran experiments from at least two of the proposals, or experiments very close to them. Not every mentee wrote a proposal. One mentee spent the early phase running experiments instead and trained many model organisms of sycophancy. Those model organisms became central to the project, and most of the other mentees went on to use them in their own experiments. It's still unclear how useful writing the proposals was.