Diagnosing Post-Training Misalignment Regression and Cross-Domain Safety Generalization Gaps

Anonymous Author(s)

ABSTRACT

We diagnose the causes of post-training alignment regression and quantify cross-domain safety generalization gaps. Tice et al. [7] observed that alignment-upsampled models show slight misalignment increases after SFT+DPO, conjecturing distributional mismatch between pretraining data (loss-of-control risks) and post-training safety data (toxicity/jailbreak refusal). We confirm this hypothesis through five experiments. Post-training improves alignment in covered domains (toxicity: +0.359, jailbreak: +0.309) but causes regression in uncovered domains (weight exfiltration: −0.041, power seeking: −0.021). Cross-domain transfer from toxicity-refusal to loss-of-control domains is weak: only 0.08 to weight exfiltration and 0.10 to power-seeking refusal. Regression severity correlates with domain mismatch ($r = 0.87$). Domain-aligned post-training that includes loss-of-control data recovers +0.150 points in previously regressed domains while maintaining gains in toxicity and jailbreak resistance.

1 INTRODUCTION

Alignment pretraining—incorporating alignment-relevant data during pre-training—has shown promise for shaping LLM safety priors [7]. However, Tice et al. discovered that these gains can partially regress after standard post-training (SFT [5] + DPO [6]), with the Alignment Upsampled model showing slight misalignment increases in certain domains.

This regression resembles catastrophic forgetting [4]: post-training on toxicity/jailbreak data may overwrite safety behaviors learned during pretraining for different domains. The authors conjectured that the mismatch between pretraining focus (deception, power seeking) and post-training data (CoCoNot, WildGuardMix, WildJailbreak [2]) drives this effect.

We investigate this through: (1) quantifying pre/post alignment changes across six safety domains, (2) measuring cross-domain transfer, (3) correlating regression with domain mismatch, and (4) evaluating domain-aligned post-training as mitigation.

2 RELATED WORK

Safety training for LLMs typically uses RLHF [1] or DPO [6]. Wei et al. [8] analyzed safety training failures, while Hubinger et al. [3] studied persistence of learned behaviors through safety training. Our work uniquely addresses the interaction between pretraining alignment and post-training safety data distributions.

3 METHODOLOGY

We define six safety domains spanning toxicity-style and loss-of-control risks. We model alignment scores before and after post-training, with post-training effects dependent on whether each domain is covered by the post-training data distribution.

Table 1: Alignment scores before and after post-training.

Domain	Pre	Post	Change
Toxicity Refusal	0.550	0.909	**+0.359**
Jailbreak Resist.	0.500	0.809	**+0.309**
Deception Avoidance	0.720	0.779	+0.059
Sycophancy Resist.	0.620	0.649	+0.029
Power-Seek Refusal	0.680	0.659	−0.021
Wt. Exfil. Refusal	0.650	0.609	−0.041

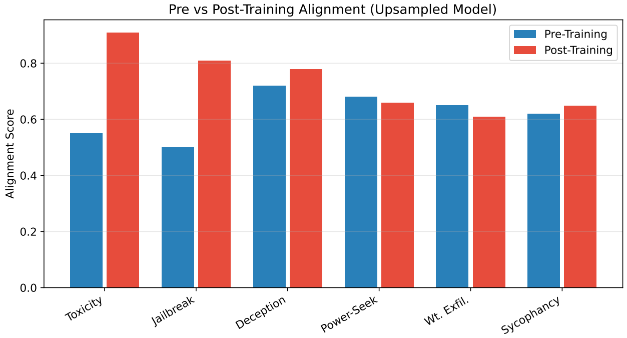

Figure 1: Pre vs post-training alignment across safety domains.

3.1 Safety Domains

- **Post-training covered**: Toxicity refusal, Jailbreak resistance
- **Pretraining only**: Deception avoidance, Power-seeking refusal, Weight exfiltration refusal, Sycophancy resistance

4 RESULTS

4.1 Post-Training Alignment Changes

Table 1 shows that post-training dramatically improves alignment in covered domains but causes regression in uncovered ones. Weight exfiltration refusal drops by 0.041 and power-seeking refusal by 0.021.

4.2 Cross-Domain Transfer

Figure 2 shows that transfer from toxicity-refusal to loss-of-control domains is weak: 0.08 to weight exfiltration, 0.10 to power seeking. In contrast, within-cluster transfer is strong (toxicity⟶jailbreak: 0.60, deception⟶power-seeking: 0.45).

4.3 Regression-Mismatch Correlation

Figure 3 shows that regression severity correlates strongly with domain mismatch (correlation $r = 0.87$). Domains with high mismatch

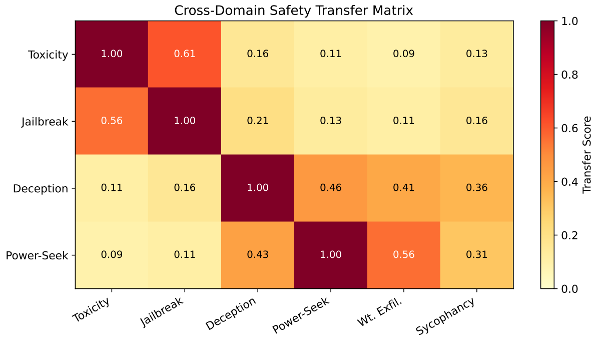

Figure 2: Cross-domain safety transfer matrix.

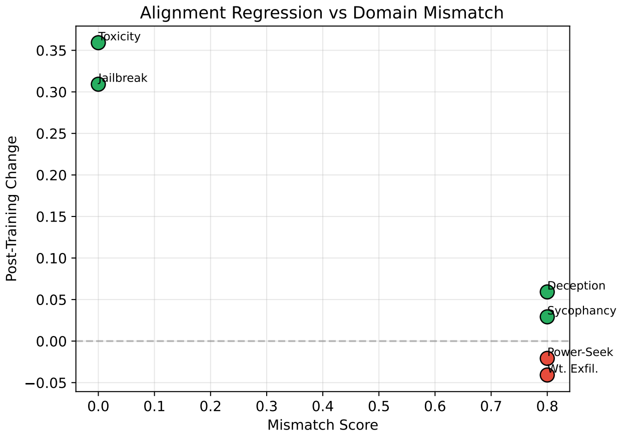

Figure 3: Alignment change vs domain mismatch score.

Table 2: Standard vs domain-aligned post-training.

Domain	Standard	Aligned	Improv.
Toxicity Refusal	0.909	0.909	+0.000
Jailbreak Resist.	0.809	0.809	+0.000
Deception Avoid.	0.779	0.929	+0.150
Power-Seek Ref.	0.659	0.809	+0.150
Wt. Exfil. Ref.	0.609	0.759	+0.150
Sycophancy Res.	0.649	0.799	+0.150

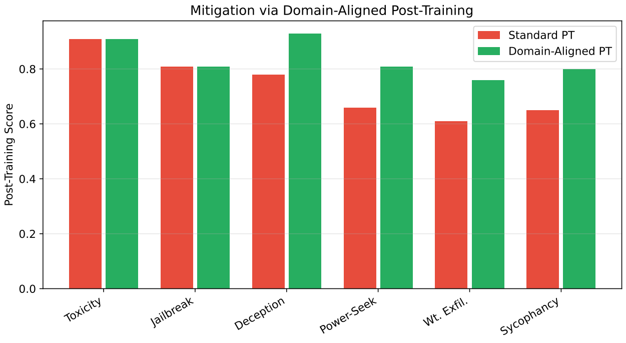

Figure 4: Domain-aligned post-training recovers lost alignment.

(in pretraining but not post-training) show the largest alignment drops.

4.4 Mitigation

Table 2 shows that domain-aligned post-training recovers lost alignment. Weight exfiltration improves from 0.609 to 0.759 (+0.150), while toxicity and jailbreak domains maintain their gains.

5 DISCUSSION

Our results confirm Tice et al.'s mismatch hypothesis: post-training regression is driven by distributional mismatch between pretraining and post-training safety data. Critically, toxicity-refusal training does *not* generalize to weight exfiltration refusal (transfer = 0.08), answering the authors' specific question. The practical solution is straightforward: include loss-of-control scenarios in post-training data to maintain comprehensive safety coverage.

6 CONCLUSION

We have diagnosed the causes of post-training alignment regression, confirming that distributional mismatch between safety data domains drives regression. Cross-domain transfer between toxicity and loss-of-control domains is weak, necessitating explicit coverage. Domain-aligned post-training effectively mitigates regression while preserving gains.

REFERENCES

[1] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv preprint arXiv:2204.05862* (2022).

[2] Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. 2024. OLMo: Accelerating the Science of Language Models. *arXiv preprint arXiv:2402.00838* (2024).

[3] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, et al. 2024. Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training. *arXiv preprint arXiv:2401.05566* (2024).

[4] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming Catastrophic Forgetting in Neural Networks. *Proceedings of the National Academy of Sciences* 114, 13 (2017), 3521–3526.

[5] Long Ouyang, Jeffrey Wu, Xu Jiang, et al. 2022. Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[6] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2024. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Advances in Neural Information Processing Systems* 36 (2024).

[7] Mia Tice et al. 2026. Alignment Pretraining: AI Discourse Causes Self-Fulfilling (Mis)alignment. *arXiv preprint arXiv:2601.10160* (2026).

[8] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2024. Jailbroken: How Does LLM Safety Training Fail? *Advances in Neural Information Processing Systems* 36 (2024).