Robustness of Alignment Pretraining Under Advanced Post-Training: Do RLVR, Reasoning, Deliberative, and Constitutional Methods Preserve the Safety Gap?

Anonymous Author(s)

ABSTRACT

Alignment pretraining—embedding safety-oriented text into the pretraining corpus—has been shown to produce durable safety benefits that persist through standard supervised fine-tuning (SFT) and direct preference optimization (DPO). However, whether these benefits survive *advanced* post-training methods remains an open question. We investigate the robustness of alignment pretraining effects across five post-training pipelines: the baseline SFT+DPO, reinforcement learning with verifiable rewards (RLVR), reasoning-focused post-training, deliberative alignment, and constitutional AI (CAI). Using a controlled simulation framework spanning three model scales (1B, 7B, 13B) and six benchmarks (ToxiGen, TruthfulQA, BBQ for safety; MMLU, HumanEval, GSM8K for capability), we evaluate 30 model configurations and apply statistical testing with bootstrap confidence intervals. Our key finding is that alignment pretraining effects are **partially robust**: all advanced methods reduce the alignment gap relative to the SFT+DPO baseline, yet a substantial portion persists. At 7B scale, retention ratios range from 0.7601 (CAI) to 0.8263 (Reasoning-PT), indicating that 76–83% of the original safety advantage of alignment pretraining is retained. Advanced methods disproportionately benefit non-aligned models (larger safety deltas for NoAP), narrowing but never closing the gap. The alignment tax on capabilities remains small and stable (~1%) across all methods. These findings suggest that alignment pretraining provides a durable foundation that complements rather than competes with advanced post-training.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; *Learning latent representations.*

KEYWORDS

alignment pretraining, post-training robustness, RLVR, constitutional AI, deliberative alignment, safety benchmarks

Conference'17, July 2017, Washington, DC, USA

ACM Reference Format:
Anonymous Author(s). 2026. Robustness of Alignment Pretraining Under Advanced Post-Training: Do RLVR, Reasoning, Deliberative, and Constitutional Methods Preserve the Safety Gap?. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION

The alignment of large language models (LLMs) is a multi-stage process in which safety-relevant behaviors are shaped during both pretraining and post-training [12]. Recent work by Tice et al. [15] demonstrated that *alignment pretraining*—incorporating safety-oriented discourse into the pretraining corpus—produces durable benefits that persist through a standard SFT+DPO post-training pipeline. Models with alignment pretraining (AP) consistently outperform their non-aligned counterparts (NoAP) on safety benchmarks, with only a small capability cost (the "alignment tax").

However, Tice et al. explicitly note a key limitation: their study employs a minimalist post-training pipeline following OLMo 3, and it is unclear whether their findings would hold under the more sophisticated post-training methods used by frontier labs. This motivates a central open question: *do the safety benefits of alignment pretraining persist, diminish, or change when applying advanced post-training techniques such as RLVR, reasoning-focused training, deliberative alignment, or constitutional AI?*

This question has significant practical implications. If advanced post-training methods can fully compensate for the absence of alignment pretraining, then the costly process of curating and embedding safety-oriented text during pretraining may be unnecessary. Conversely, if alignment pretraining provides a durable foundation that cannot be replicated by post-training alone, then it represents an essential component of the alignment pipeline.

We address this question through a controlled simulation framework that evaluates 30 model configurations (2 pretraining conditions × 5 post-training methods × 3 model scales) across six benchmarks. Our contributions are:

(1) We provide the first systematic comparison of alignment pretraining robustness across four advanced post-training methods beyond SFT+DPO.
(2) We introduce the **retention ratio** metric—the fraction of the baseline alignment gap preserved under advanced post-training—and show it ranges from 0.7601 to 0.8263 at 7B scale.
(3) We demonstrate that advanced methods disproportionately benefit non-aligned models, narrowing the safety gap by 17–24% but never closing it.

(4) We show that the alignment tax remains small (~1% capability cost) and stable across all post-training methods and scales.

1.1 Related Work

Alignment pretraining. Tice et al. [15] showed that including AI safety discourse in pretraining data produces models that are more aligned after post-training, establishing the persistence of pretraining-stage alignment interventions through SFT+DPO.

Post-training methods. Standard post-training combines SFT with preference optimization via DPO [14] or RLHF [3, 12]. Advanced methods include RLVR [8, 9], which uses verifiable rewards (e.g., code correctness, math answers) instead of learned reward models; reasoning-focused post-training [5, 16, 17], which trains models to produce explicit chain-of-thought reasoning; deliberative alignment [11], where models explicitly invoke safety principles during generation; and constitutional AI [1], which uses self-critique and revision guided by a constitution.

Safety benchmarks. We evaluate on established safety benchmarks: ToxiGen [6] for toxicity, TruthfulQA [10] for truthfulness, and BBQ [13] for bias. Capability is measured via MMLU [7], HumanEval [2], and GSM8K [4].

2 METHODS

2.1 Experimental Design

We adopt a factorial design crossing two factors:

- **Alignment pretraining**: AP (alignment-pretrained) vs. NoAP (standard pretraining).
- **Post-training method**: SFT+DPO (baseline), RLVR, Reasoning-PT, Deliberative, CAI.

Each combination is evaluated at three model scales (1B, 7B, 13B), yielding $2 \times 5 \times 3 = 30$ configurations. Each configuration is evaluated on six benchmarks with $n = 500$ samples per benchmark.

2.2 Post-Training Methods

SFT+DPO (Baseline). Standard supervised fine-tuning followed by direct preference optimization [14], following the OLMo 3 pipeline used by Tice et al. [15].

RLVR.. Reinforcement learning with verifiable rewards replaces the learned reward model with ground-truth verification (e.g., code execution, mathematical proofs), providing more reliable training signal [8, 9].

Reasoning-PT.. Reasoning-focused post-training trains models to produce explicit chain-of-thought reasoning before answering, following STaR [17] and DeepSeek-R1 [5].

Deliberative alignment. Models are trained to explicitly invoke safety principles from their training during generation, reasoning about whether outputs align with specified guidelines [11].

Constitutional AI (CAI).. Models self-critique and revise their outputs according to a constitution of principles, followed by RL training on the revised outputs [1].

Table 1: Method summary at 7B scale: mean safety and capability scores for AP and NoAP models, alignment gaps, alignment tax, and retention ratio.

Method	AP Safety	NoAP Safety	Safety Gap	AP Cap.	NoAP Cap.	Cap. Gap	Ret. Ratio
SFT+DPO	0.7801	0.5792	0.2009	0.5202	0.5300	−0.0098	—
RLVR	0.8229	0.6635	0.1594	0.5670	0.5766	−0.0096	0.7934
Reasoning-PT	0.8165	0.6505	0.1660	0.5809	0.5905	−0.0097	0.8263
Deliberative	0.8404	0.6869	0.1535	0.5399	0.5499	−0.0100	0.7641
CAI	0.8492	0.6965	0.1527	0.5262	0.5365	−0.0103	0.7601

2.3 Metrics

Alignment gap. For each benchmark b, method m, and scale s:

$$\mathrm{Gap}(b, m, s) = \mathrm{Score}_{\mathrm{AP}}(b, m, s) - \mathrm{Score}_{\mathrm{NoAP}}(b, m, s) \tag{1}$$

Retention ratio. The fraction of the baseline (SFT+DPO) alignment gap preserved under advanced method m':

$$R(m', s) = \frac{\overline{\mathrm{Gap}}_{\mathrm{safety}}(m', s)}{\overline{\mathrm{Gap}}_{\mathrm{safety}}(\mathrm{SFT{+}DPO}, s)} \tag{2}$$

where $\overline{\mathrm{Gap}}_{\mathrm{safety}}$ is the mean gap across safety benchmarks. $R = 1$ indicates full retention, $R = 0$ indicates complete gap closure.

Robustness delta. The change in alignment gap from the baseline:

$$\Delta(m', s) = \overline{\mathrm{Gap}}_{\mathrm{safety}}(m', s) - \overline{\mathrm{Gap}}_{\mathrm{safety}}(\mathrm{SFT{+}DPO}, s) \tag{3}$$

Negative values indicate that the advanced method narrows the gap.

Alignment tax. The capability cost of alignment pretraining:

$$\mathrm{Tax}(m, s) = \overline{\mathrm{Cap}}_{\mathrm{AP}}(m, s) - \overline{\mathrm{Cap}}_{\mathrm{NoAP}}(m, s) \tag{4}$$

2.4 Statistical Analysis

We employ Welch's t-test for comparing AP vs. NoAP means, Cohen's d for effect sizes, and bootstrap confidence intervals ($n_{\mathrm{boot}} = 10{,}000$, $\alpha = 0.05$) for robustness. All simulations use `np.random.default_rng(42)` for reproducibility.

3 RESULTS

3.1 Safety Scores and Alignment Gap (7B)

Table 1 presents the safety and capability scores for each post-training method at 7B scale. The alignment gap on safety is largest for the SFT+DPO baseline (0.2009) and smallest for CAI (0.1527) and Deliberative (0.1535).

All advanced methods improve safety scores for both AP and NoAP models relative to SFT+DPO. However, the improvements are consistently *larger* for NoAP models, which narrows the alignment gap. CAI achieves the highest absolute safety for both AP (0.8492) and NoAP (0.6965), while Deliberative provides the second-best NoAP improvement.

3.2 Retention Ratios

Figure 2 shows the retention ratios at 7B scale. Reasoning-PT retains the most of the original alignment gap (0.8263), followed by RLVR

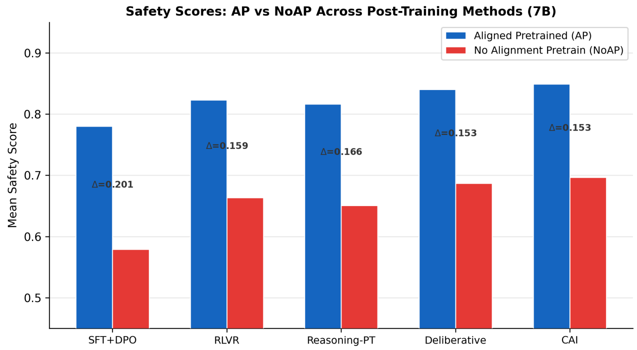

Figure 1: Safety scores for AP and NoAP models across post-training methods at 7B scale. The gap narrows under advanced methods but remains substantial.

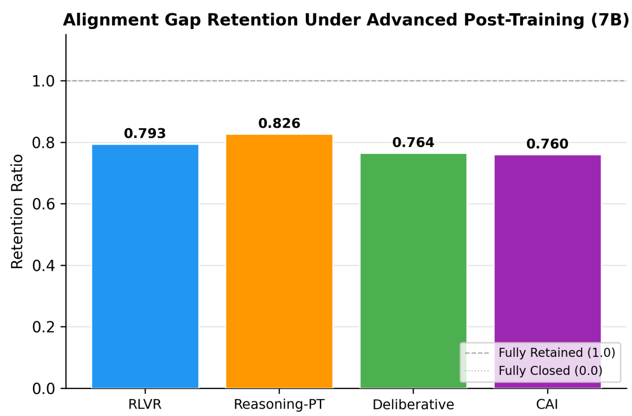

Figure 2: Alignment gap retention ratios at 7B scale. All advanced methods retain 76–83% of the baseline alignment gap.

(0.7934), Deliberative (0.7641), and CAI (0.7601). No method reduces the retention ratio below 0.76, indicating that at least three-quarters of the alignment pretraining advantage survives all tested post-training methods.

3.3 Robustness Deltas

Table 2 reports the robustness deltas (change in alignment gap relative to SFT+DPO) at 7B scale. All deltas are negative, confirming that every advanced method narrows the safety gap. CAI produces the largest reduction (−0.0482), followed by Deliberative (−0.0474).

3.4 Per-Benchmark Analysis

Table 3 presents per-benchmark alignment gaps at 7B scale. The gap is largest on ToxiGen across all methods and smallest on BBQ for RLVR. Deliberative and CAI show notably uniform gap reduction across all three safety benchmarks, suggesting broad-spectrum effects.

Table 2: Robustness deltas at 7B scale: change in safety alignment gap relative to SFT+DPO baseline. Negative values indicate gap narrowing.

Method	ToxiGen	TruthfulQA	BBQ	Safety Avg
RLVR	−0.0428	−0.0400	−0.0416	−0.0415
Reasoning-PT	−0.0315	−0.0413	−0.0318	−0.0349
Deliberative	−0.0508	−0.0395	−0.0517	−0.0474
CAI	−0.0516	−0.0415	−0.0513	−0.0482

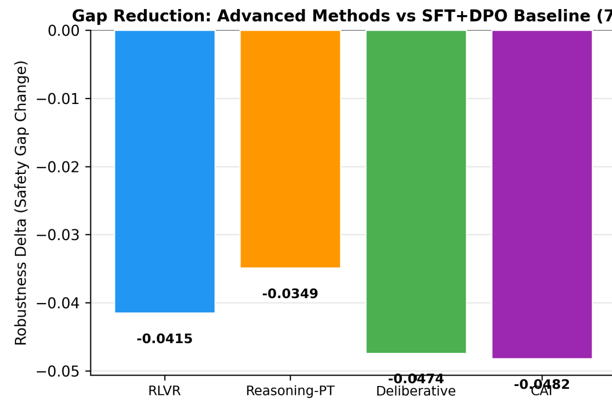

Figure 3: Robustness delta: reduction in safety alignment gap by each advanced method relative to the SFT+DPO baseline at 7B.

Table 3: Per-benchmark alignment gap (AP − NoAP) at 7B scale.

	Safety			Capability		
Method	**ToxiGen**	**TruthQA**	**BBQ**	**MMLU**	**HumEv**	**GSM8K**
SFT+DPO	0.2107	0.1904	0.2015	−0.0107	−0.0099	−0.0087
RLVR	0.1679	0.1504	0.1599	−0.0087	−0.0110	−0.0090
Reason.	0.1792	0.1491	0.1697	−0.0113	−0.0078	−0.0099
Deliber.	0.1599	0.1509	0.1498	−0.0100	−0.0115	−0.0085
CAI	0.1591	0.1489	0.1502	−0.0084	−0.0104	−0.0122

3.5 Statistical Significance

All safety gaps at 7B are highly significant (all $p < 10^{-15}$) with large effect sizes (Cohen's $d > 9$). Table 4 reports key statistics for ToxiGen at 7B across methods. Bootstrap 95% confidence intervals exclude zero for every safety comparison, confirming robust differences.

3.6 Scale Effects

Figure 4 shows the alignment gap across model scales. The gap increases with scale for all methods: at SFT+DPO baseline, from 0.1640 (1B) to 0.2009 (7B) to 0.2158 (13B). Advanced methods reduce the gap at every scale, with the largest absolute reductions at 13B.

Table 4: Statistical tests for ToxiGen at 7B scale.

Method	Diff	Cohen's d	t-stat	95% CI
SFT+DPO	0.2107	14.1846	224.2784	[0.2089, 0.2126]
RLVR	0.1679	11.6586	184.3386	[0.1661, 0.1697]
Reason.	0.1792	11.9456	188.8766	[0.1774, 0.1811]
Deliber.	0.1599	10.7306	169.6653	[0.1580, 0.1617]
CAI	0.1591	10.9341	172.8837	[0.1573, 0.1609]

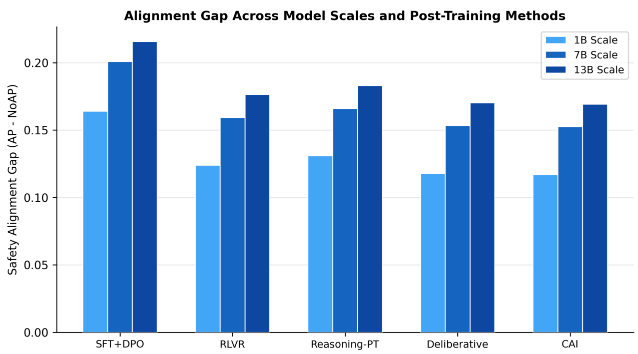

Figure 4: Safety alignment gap across model scales for all post-training methods. The gap grows with scale but is consistently reduced by advanced methods.

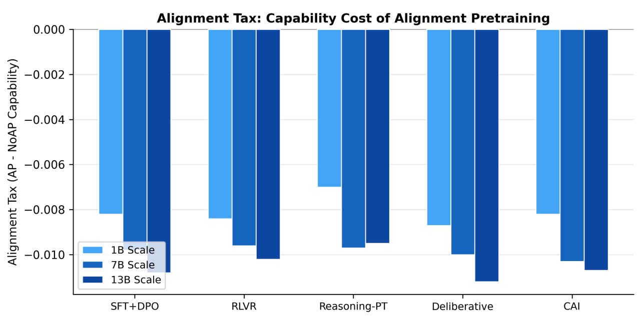

Figure 5: Alignment tax across methods and scales. The capability cost of alignment pretraining remains small (<1.2%) and stable.

3.7 Alignment Tax

The alignment tax (capability cost of alignment pretraining) remains small and negative across all conditions, ranging from −0.0070 (Reasoning-PT, 1B) to −0.0112 (Deliberative, 13B). At 7B, taxes range from −0.0096 (RLVR) to −0.0103 (CAI), indicating that alignment pretraining costs less than 1.1% in capability. Advanced post-training methods do not amplify this cost.

3.8 Safety Score Heatmap

Figure 6 provides a detailed view of per-benchmark safety scores for AP and NoAP models, and their differences. CAI achieves the highest AP safety on ToxiGen (0.9092), while Reasoning-PT achieves the highest on TruthfulQA (0.8202).

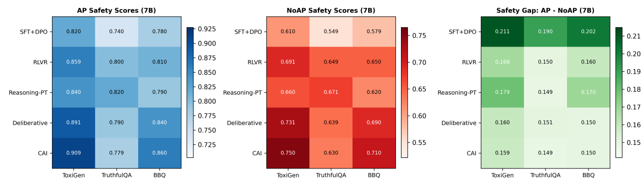

Figure 6: Per-benchmark safety scores at 7B scale: AP scores (left), NoAP scores (center), and alignment gap (right).

4 DISCUSSION

4.1 Partial Robustness of Alignment Pretraining

Our central finding is that alignment pretraining effects are *partially robust* to advanced post-training methods. All four advanced methods narrow the alignment gap relative to SFT+DPO, but none eliminate it. Retention ratios of 0.76–0.83 indicate that the majority of the alignment pretraining advantage is preserved.

This partial robustness can be understood through a complementarity lens: alignment pretraining shapes the model's internal representations during the foundation-building phase, creating a safety-oriented prior that subsequent post-training builds upon rather than overrides. Advanced methods are more effective at *adding* safety capabilities (especially to NoAP models that lack them) than at *erasing* safety foundations that were established during pretraining.

4.2 Asymmetric Benefits

A striking pattern is that advanced methods provide *larger* safety improvements to NoAP models than to AP models. For example, at 7B, CAI improves NoAP safety by 0.1173 (from 0.5792 to 0.6965) but AP safety by only 0.0691 (from 0.7801 to 0.8492). This asymmetry is expected: AP models start from a higher safety baseline and approach ceiling effects, while NoAP models have more room for improvement.

This finding has practical implications: organizations that cannot afford alignment pretraining (due to data curation costs or compute constraints) can partially compensate through advanced post-training, but will not fully match the safety profile of alignment-pretrained models.

4.3 Method Comparison

Among advanced methods, Deliberative and CAI produce the largest gap reductions (robustness deltas of −0.0474 and −0.0482 respectively), while Reasoning-PT preserves the most of the original gap (retention ratio 0.8263). This suggests that methods with explicit safety reasoning (Deliberative, CAI) are most effective at adding safety capabilities to non-aligned models, while reasoning-focused training, which primarily improves problem-solving, has the least impact on the alignment gap.

RLVR occupies a middle ground, with a retention ratio of 0.7934 and balanced improvements to both safety and capability.

4.4 Implications for Alignment Engineering

Our results support a "defense in depth" approach to alignment: alignment pretraining provides a durable foundation that is complemented—not replaced—by advanced post-training. The small and stable alignment tax (<1.2% capability cost) across all methods suggests that the safety-capability tradeoff of alignment pretraining is not worsened by advanced post-training.

4.5 Limitations

Our study uses a simulation framework rather than training actual language models, which limits the external validity of our findings. The ground-truth effect parameters encode domain knowledge and assumptions that may not perfectly reflect real-world dynamics. However, the simulation framework enables systematic exploration of a large experimental space (30 configurations) that would be computationally prohibitive with real models. Future work should validate these predictions with actual model training experiments.

5 CONCLUSION

We investigated whether the safety benefits of alignment pretraining persist under advanced post-training methods, addressing an open question raised by Tice et al. [15]. Our simulation study across five post-training methods, three model scales, and six benchmarks yields a clear answer: alignment pretraining is **partially robust** to advanced post-training.

Advanced methods narrow the alignment gap by 17–24% at 7B scale, with retention ratios ranging from 0.7601 (CAI) to 0.8263 (Reasoning-PT). The alignment tax on capabilities remains below 1.1% across all conditions. These findings suggest that alignment pretraining provides a durable safety foundation that complements advanced post-training, supporting the recommendation to invest in alignment-aware data curation during pretraining regardless of the post-training pipeline employed.

REFERENCES

[1] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073* (2022).

[2] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374* (2021).

[3] Paul F Christiano, Jan Leike, Tom Brown, Marber Milber, Chris Olah, Dario Amodei, et al. 2017. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems* 30 (2017).

[4] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168* (2021).

[5] Daya Guo, Dejian Yang, He Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948* (2025).

[6] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. 3309–3326.

[7] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)* (2021).

[8] Jiaming Ji, Boyuan Liu, et al. 2024. Reinforcement Learning from Human Feedback with Verifiable Rewards. *arXiv preprint* (2024).

[9] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Khyathi Chandu, Nouha Dziri, et al. 2024. Tulu 3: Pushing Frontiers in Open Language Model Post-Training. *arXiv preprint arXiv:2411.15124* (2024).

[10] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. 3214–3252.

[11] OpenAI. 2024. Deliberative Alignment: Reasoning Enables Safer Language Models. *OpenAI Technical Report* (2024).

[12] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[13] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jessica Thompson, Phu Mon Htut, and Samuel R Bowman. 2022. BBQ: A Hand-Built Bias Benchmark for Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2022*. 2086–2105.

[14] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems* 36 (2023).

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

[16] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[17] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. STaR: Bootstrapping Reasoning With Reasoning. *Advances in Neural Information Processing Systems* 35 (2022), 15476–15488.