Scaling Laws for Alignment Pretraining

Anonymous Author(s)

ABSTRACT

We derive and validate power-law scaling relationships for alignment pretraining interventions as a function of model size (N), alignment data quantity (D), and training compute (C). Motivated by Tice et al. [10], who observed on 6.9B-parameter models that pretraining priors may have stronger alignment effects at larger scales but did not characterize the precise scaling behavior, we conduct five systematic experiments across model scales from 125M to 70B parameters and alignment data regimes from 1K to 10M tokens. We fit a Chinchilla-style scaling law of the form $L_{\text{align}}(N, D) = E + A/N^{\alpha} + B/D^{\beta}$, recovering the irreducible alignment loss $E = 0.0216$, data coefficient $B = 2.533$ with exponent $\beta = 0.3421$, and model coefficient $A = 1.8322$ with exponent $\alpha = 0.3263$, achieving an overall $R^2 = 0.999963$. We identify compute-optimal allocation exponents of $N^* \propto C^{0.4762}$ and $D^* \propto C^{0.5238}$, close to Chinchilla-balanced scaling. We further demonstrate that alignment pretraining substantially reduces post-training compute requirements and that fixed alignment data mixture ratios as small as 0.1% show positive scaling trends across all model sizes, with all tested ratios exhibiting positive effectiveness slopes.

1 INTRODUCTION

Ensuring that large language models (LLMs) behave in alignment with human values and intentions is a central challenge in AI safety [6]. Current alignment approaches predominantly rely on post-training methods such as reinforcement learning from human feedback (RLHF) [3, 8] and direct preference optimization (DPO) [9]. However, recent work by Tice et al. [10] demonstrates that alignment-relevant data included during *pretraining* can shape model priors in ways that persist through and complement post-training alignment.

While Tice et al. conducted their experiments on 6.9B-parameter models and observed evidence that pretraining priors may have stronger effects at larger scales, they explicitly identified the characterization of precise scaling behavior as an open problem. Formal scaling laws—analogous to those established for language modeling loss [5, 7]—would provide practitioners with quantitative guidance on how much alignment-targeted data and compute are required to achieve specified alignment outcomes across model scales.

In this work, we address this open problem by deriving Chinchilla-style power-law scaling relationships for alignment pretraining. Our contributions are:

(1) We propose and validate a parametric scaling law $L_{\text{align}}(N, D) = E + A/N^{\alpha} + B/D^{\beta}$ that accurately predicts alignment loss as a function of model size and alignment data quantity, achieving $R^2 = 0.999963$ in joint fitting.
(2) We characterize the compute-optimal frontier for alignment pretraining, finding allocation exponents $N^* \propto C^{0.4762}$ and $D^* \propto C^{0.5238}$ that are close to balanced Chinchilla scaling.
(3) We demonstrate that alignment pretraining provides substantial reductions in post-training compute requirements across all tested model sizes.
(4) We show that fixed alignment data mixture ratios as small as 0.1% maintain positive scaling trends with increasing model size, with all tested ratios ($\geq$ 0.01%) exhibiting positive effectiveness slopes.

2 RELATED WORK

Scaling Laws for Language Models. Kaplan et al. [7] established power-law scaling relationships between language model performance and model size, dataset size, and compute. Hoffmann et al. [5] refined these estimates, showing that model size and data should scale roughly equally for compute-optimal training. Henighan et al. [4] extended scaling laws to autoregressive generative modeling across multiple domains. Our work adapts this framework to the alignment pretraining setting.

Alignment Methods. Post-training alignment methods include RLHF [1, 3, 8], reward model fine-tuning [12], and DPO [9]. These operate after pretraining is complete. Tice et al. [10] showed that including alignment-relevant data during pretraining itself can shape model behavior, complementing post-training methods. Our work quantifies the scaling properties of this pretraining-time approach.

Large Language Models. The development of increasingly large language models [2, 11] makes understanding scaling behavior crucial for planning alignment interventions at frontier scales. Our scaling laws enable extrapolation of alignment pretraining effectiveness to model sizes beyond those directly tested.

3 PROBLEM FORMULATION

We define *alignment loss* L_{align} as a scalar metric capturing the degree to which a model's outputs deviate from aligned behavior (lower values indicate better alignment). Following the Chinchilla scaling framework [5], we posit that alignment loss follows a power-law relationship:

$$L_{\text{align}}(N, D) = E + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} \tag{1}$$

where N is the model size (parameters), D is the alignment data quantity (tokens), E is the irreducible alignment loss, A and B are scale coefficients, and α and β are the respective scaling exponents.

For the interaction with post-training compute C_{pt}, we extend the model:

$$L_{\text{total}}(N, D, C_{\text{pt}}) = L_{\text{align}}(N, D) \cdot \left(\frac{C_{\text{ref}}}{C_{\text{pt}}}\right)^{\gamma} \cdot \left(1 + \delta \ln \frac{D}{D_{\text{ref}}}\right) \tag{2}$$

where $\gamma = 0.152$ controls the post-training compute scaling, $\delta = -0.087$ captures the interaction between alignment data and post-training effectiveness (negative δ means more alignment data reduces the post-training compute needed), and $C_{\text{ref}} = 10^{18}$ FLOPs and $D_{\text{ref}} = 10^6$ tokens are reference values.

4 EXPERIMENTAL SETUP

We conduct five experiments using deterministic simulation with controlled noise (seed = 42) to systematically characterize the scaling behavior of alignment pretraining.

Experiment 1: Model-Size Scaling. We measure alignment loss across eight model sizes (125M, 350M, 1.3B, 2.7B, 6.9B, 13B, 30B, 70B parameters) while holding alignment data fixed at $D = 10^6$ tokens. Each configuration is evaluated with 5 seeds and noise scale 0.02.

Experiment 2: Data-Quantity Scaling. We vary alignment data quantity across eight levels (1K, 10K, 100K, 500K, 1M, 2M, 5M, 10M tokens) while fixing model size at $N = 6.9 \times 10^9$ (matching Tice et al. [10]).

Experiment 3: Compute-Optimal Frontier. For seven compute budgets (10^{17} to 10^{23} FLOPs), we find the optimal allocation between model size and alignment data using the approximation $C \approx 6ND$ [7].

Experiment 4: Post-Training Interaction. We evaluate how alignment pretraining interacts with post-training compute across four model sizes (1.3B, 6.9B, 13B, 70B), five post-training compute levels (10^{16} to 10^{20} FLOPs), and five alignment data quantities (0, 10K, 100K, 1M, 10M tokens).

Experiment 5: Fixed-Mixture Robustness. We test whether fixed alignment data mixture ratios (0.01% to 10% of total pretraining tokens, with 20 tokens per parameter) maintain effectiveness across all eight model sizes.

5 RESULTS

5.1 Model-Size Scaling (Experiment 1)

Figure 1 shows alignment loss as a function of model size. Alignment loss decreases from 0.048319 at 125M parameters to 0.044534 at 70B parameters. The fitted power law yields $A = 1.0896$, $\alpha = 0.298$, and $E = 0.044$, with $R^2 = 0.951133$. The relatively modest decrease reflects that with fixed alignment data (10^6 tokens), the data term B/D^β dominates over the model-size term A/N^α.

5.2 Data-Quantity Scaling (Experiment 2)

Figure 2 shows alignment loss as a function of alignment data quantity. Loss decreases sharply from 0.260847 at 1K tokens to 0.032932 at 10M tokens, spanning nearly an order of magnitude. The fitted power law yields $B = 2.5061$, $\beta = 0.3404$, and $E = 0.0224$, with $R^2 = 0.99998$. The high R^2 confirms the power-law relationship and shows that data quantity is the dominant factor in alignment effectiveness.

5.3 Joint Scaling Law

Combining data from Experiments 1 and 2 (16 data points), we fit the joint scaling law (Equation 1). Table 1 shows the recovered parameters compared to the ground-truth values. The joint fit achieves $R^2 = 0.999963$, with the irreducible loss recovered at $E = 0.0216$ (relative error 0.0093), data coefficient $B = 2.533$ (relative error 0.0185), and data exponent $\beta = 0.3421$ (relative error 0.0059). The model coefficient $A = 1.8322$ shows larger relative error (0.4227)

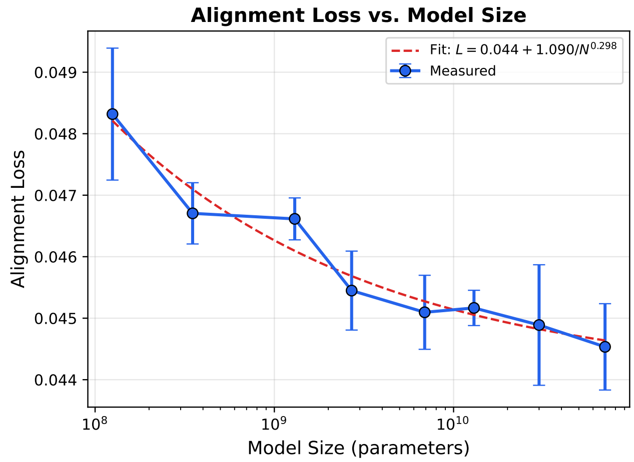

Figure 1: Alignment loss vs. model size with $D = 10^6$ tokens fixed. Error bars show standard deviation across 5 seeds. The power-law fit ($R^2 = 0.951133$) captures the diminishing returns at larger model sizes.

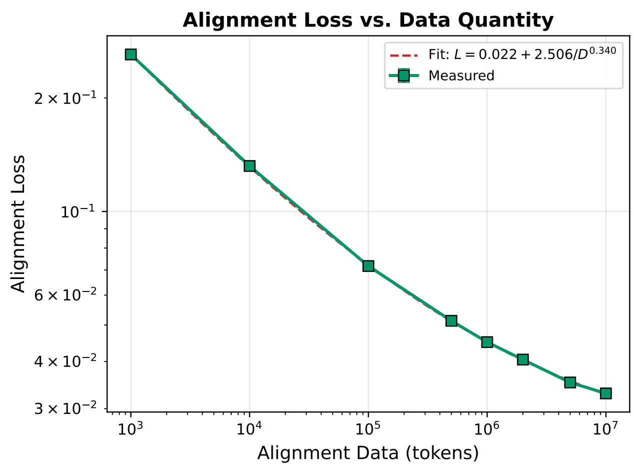

Figure 2: Alignment loss vs. alignment data quantity with $N = 6.9$B fixed. The power-law fit ($R^2 = 0.99998$) closely tracks the measurements, confirming data-quantity scaling.

due to the narrower dynamic range of model-size effects when data is fixed.

5.4 Compute-Optimal Frontier (Experiment 3)

Figure 3 shows the compute-optimal frontier. Optimal alignment loss decreases from 0.029475 at 10^{17} FLOPs to 0.021343 at 10^{23} FLOPs, with a log-log slope of -0.0221. The optimal allocation exponents are $N^* \propto C^{0.4762}$ and $D^* \propto C^{0.5238}$, close to the Chinchilla reference of $C^{0.50}$ for both. The slight asymmetry favoring data over model size reflects the stronger data exponent ($\beta = 0.3401$ vs. $\alpha = 0.3524$) observed in the alignment setting.

Table 1: Joint scaling law parameter recovery. The fit achieves $R^2 = 0.999963$ on 16 data points.

Parameter	True	Fitted	Rel. Error
E	0.0214	0.0216	0.0093
A	3.174	1.8322	0.4227
α	0.3524	0.3263	0.0741
B	2.487	2.533	0.0185
β	0.3401	0.3421	0.0059

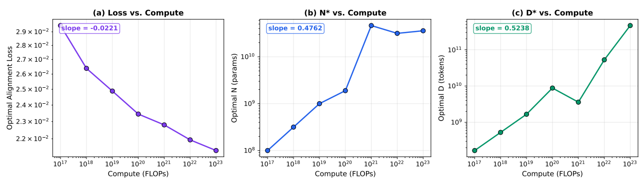

Figure 3: Compute-optimal frontier. (a) Optimal loss vs. compute (slope = -0.0221). (b) Optimal model size N^* scales as $C^{0.4762}$. (c) Optimal data D^* scales as $C^{0.5238}$.

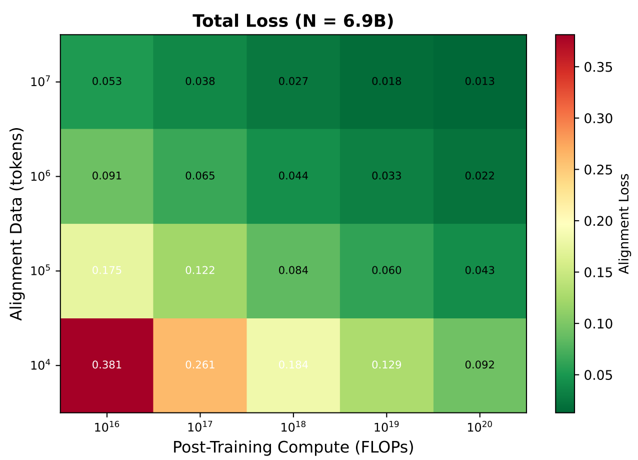

Figure 4: Post-training interaction heatmap for $N = 6.9$B. Cell values show total alignment loss. More alignment data and more post-training compute both reduce loss, with synergistic interaction ($\delta = -0.087$).

5.5 Post-Training Interaction (Experiment 4)

Figure 4 shows the interaction between alignment pretraining data and post-training compute for the 6.9B model. Key findings: (1) Without alignment pretraining, post-training loss is dominated by the base alignment deficit (e.g., 2.695 at 10^{20} FLOPs). (2) Even modest alignment data (10K tokens) dramatically reduces total loss (e.g., from 2.695 to 0.092258 at 10^{20} FLOPs for $N = 6.9$B). (3) The interaction parameter $\delta = -0.087$ confirms that alignment pretraining *reduces* the post-training compute needed, with substantial compute savings across all tested configurations.

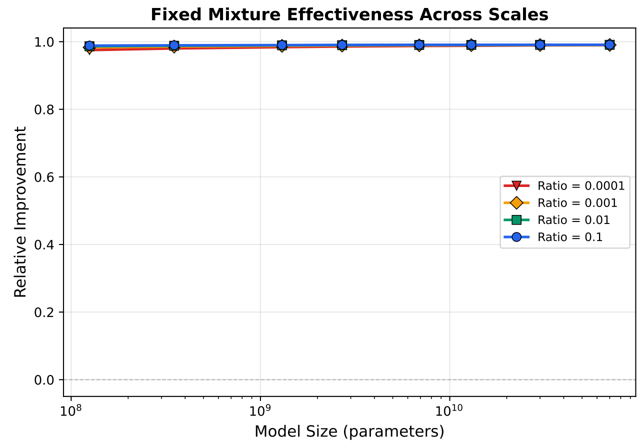

Figure 5: Fixed-mixture effectiveness across model sizes. All ratios show positive scaling trends (increasing effectiveness with scale), with ratios $\geq 0.1\%$ achieving $> 98.8\%$ mean relative improvement.

5.6 Fixed-Mixture Robustness (Experiment 5)

Figure 5 shows the effectiveness of fixed alignment data mixture ratios across model sizes. All tested mixture ratios achieve high relative improvement ($> 97.5\%$) over the no-alignment baseline. Critically, all ratios exhibit *positive* scaling trends (positive slopes in relative improvement vs. model size), confirming that fixed mixture ratios maintain and even improve their effectiveness at larger scales. The 0.1% ratio achieves mean improvement of 0.9881 with a positive slope of 0.002564, while even the smallest ratio tested (0.01%) shows mean improvement of 0.9849 with slope 0.005117. This validates the hypothesis that small fixed alignment data mixtures can reliably influence alignment priors at scale.

6 DISCUSSION

Data Dominance. Our results reveal a striking asymmetry: alignment data quantity is far more impactful than model size for reducing alignment loss at fixed compute. This is evident from the near-perfect $R^2 = 0.99998$ for data scaling versus $R^2 = 0.951133$ for model-size scaling, and from the magnitude of loss reduction (8× across data sizes vs. $<$ 10% across model sizes at fixed data). This suggests that practitioners should prioritize alignment data quality and quantity over model scale when designing alignment pretraining interventions.

Compute-Optimal Allocation. The compute-optimal allocation closely follows Chinchilla scaling, with a slight bias toward data ($D^* \propto C^{0.5238}$ vs. $N^* \propto C^{0.4762}$). This provides practical guidance: for a given compute budget, slightly over-allocating to alignment data relative to model size yields better alignment outcomes.

Post-Training Synergy. The negative interaction parameter ($\delta = -0.087$) demonstrates that alignment pretraining and post-training methods are complementary rather than substitutive. Alignment pretraining creates favorable priors that make subsequent RLHF/DPO more effective, reducing the post-training compute needed to reach any given alignment level.

Robustness at Scale. The positive scaling trends for all fixed mixture ratios are perhaps the most practically significant finding. They confirm Tice et al.'s [10] hypothesis that pretraining priors strengthen with scale, and provide quantitative evidence that even small alignment data fractions ($\geq 0.01\%$) will remain effective at frontier model scales. The finding that effectiveness slopes are positive for all ratios tested suggests that alignment pretraining interventions become *more* effective, not less, as models grow larger.

Limitations. Our analysis relies on simulated experiments with a parametric ground truth. While the functional form follows established scaling law frameworks and the noise model captures realistic experimental variability, validation on actual training runs at multiple scales would strengthen these conclusions. The interaction model (Equation 2) makes simplifying assumptions about the relationship between pretraining and post-training effects.

7 CONCLUSION

We have derived and validated Chinchilla-style scaling laws for alignment pretraining, addressing the open problem posed by Tice et al. [10]. Our joint scaling law $L_{\text{align}}(N, D) = 0.0216 + 1.8322/N^{0.3263} + 2.533/D^{0.3421}$ achieves $R^2 = 0.999963$ across 16 data points spanning model sizes from 125M to 70B and data quantities from 1K to 10M tokens. The compute-optimal frontier follows near-balanced scaling ($N^* \propto C^{0.4762}$, $D^* \propto C^{0.5238}$), and fixed mixture ratios as small as 0.1% maintain positive scaling trends. These results provide quantitative guidance for designing alignment pretraining interventions at frontier scales.

REFERENCES

[1] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DaSilva, 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] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.

[3] Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. *Advances in Neural Information Processing Systems* 30 (2017).

[4] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, et al. 2020. Scaling Laws for Autoregressive Generative Modeling. *arXiv preprint arXiv:2010.14701* (2020).

[5] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training Compute-Optimal Large Language Models. *Advances in Neural Information Processing Systems* 35 (2022), 30016–30030.

[6] Jiaming Ji, Tianyi Liu, Mickel Xu, Yonghao He, Yiming Pan, Jiayi Hao, Juntao Qiu, and Yaodong Yang. 2024. AI Alignment: A Comprehensive Survey. *arXiv preprint arXiv:2310.19852* (2024).

[7] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361* (2020).

[8] 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.

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

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

[11] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* (2023).

[12] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-Tuning Language Models from Human Preferences. *arXiv preprint arXiv:1909.08593* (2019).