

---

# Scaling AI Safety: Empirical Measurements of Technique Effectiveness Across Model Sizes

---

Kyuhee Kim  
EPFL

Ines Altemir Marinas  
EPFL

Arthur Wuhrmann  
EPFL

With  
Apart Research

## Abstract

This study represents a first attempt to empirically measure how AI safety interventions scale with model size—a critical yet overlooked factor in AI capability forecasting. We analyze published results and conduct additional evaluations to measure performance improvements from safety techniques including RLHF, DPO, and unlearning methods on benchmarks such as HarmBench, WMDP, and ToxiGen. Using linear regression on relative performance improvements across model scales from 1B to 236B parameters, we tried to investigate whether safety gains increase, decrease, or remain constant as models grow larger. Our analysis reveals that most existing safety research tests interventions at only one or two model scales, limiting our ability to draw robust conclusions about scaling patterns. The sparse data yields low statistical significance, but observed trends suggest safety interventions may exhibit different scaling behaviors across techniques and benchmarks. This work establishes a methodology for measuring safety scaling and demonstrates the need for systematic multi-scale experiments in safety research. We provide initial empirical data that can inform AI timeline forecasting models by quantifying how safety requirements might evolve with model scale.

*Keywords: scaling laws, AI safety benchmarks, empirical analysis,*

# 1. Introduction

The rapid advancement of large language models has prompted extensive research into forecasting AI development timelines, yet current methodologies primarily focus on capability scaling laws (Kaplan et al., 2020). A critical but underexplored dimension is how AI safety interventions scale alongside these capabilities.

Recent work by Ren et al. (2024) revealed that many AI safety benchmarks correlate strongly with general capabilities, raising questions about whether observed safety improvements reflect genuine progress or merely capability advancement. This finding suggests our understanding of when AI systems can be safely deployed may be fundamentally miscalibrated.

This study addresses a crucial empirical gap: **how do AI safety intervention techniques—such as RLHF and Unlearning—scale in effectiveness as models increase in size?** While capability scaling follows well-established power laws, the scaling properties of safety interventions remain largely unmeasured, despite their critical importance for deployment decisions and timeline predictions.

We analyze safety technique effectiveness across model scales by examining performance gaps between base and safety-enhanced models. Using linear regression on benchmark improvements, we tried to investigate whether safety gains remain constant, increase, or diminish as models grow. By focusing on decorrelated safety metrics that measure properties distinct from general capabilities, we isolate genuine safety improvements.

Our analysis encompasses 11 recent papers covering 286 model-intervention-benchmark combinations across scales from 1B to 236B parameters. Our findings provide empirical data on how safety intervention effectiveness changes with model scale, offering crucial inputs for AI timeline forecasting. Understanding these scaling dynamics is essential for predicting not just when transformative AI will emerge, but when it can be safely deployed.

## 2. Methodology

### 2.1 Data Collection

Given time constraints in the hackathon setting, we prioritized gathering data from recent publications on AI safety techniques. We systematically reviewed papers from 2023-2024 that reported safety benchmark performance before and after interventions.

We analyzed the following safety intervention techniques:

- Supervised Fine-Tuning (SFT) with safety data
- Direct Preference Optimization (DPO)
- Reinforcement Learning from Verifiable Rewards (RLVR)

- In-Context Learning (ICL) with safety prompts
- Unlearning approaches, particularly Representation Misdirection for Unlearning (RMU)

Our analysis focused on these primary safety benchmarks:

- HarmBench: Measuring resistance to harmful requests
- WMDP: Assessing dangerous knowledge suppression
- ToxiGen: Evaluating toxic content generation
- TruthfulQA: Testing truthfulness and misconception avoidance

To address gaps in published data, we conducted our own evaluations where necessary. When papers reported results at only one model scale, we ran additional experiments using publicly available model checkpoints to obtain multi-scale data points. This combination of literature analysis and original evaluation provided a more complete dataset for scaling analysis.

Final dataset: 11 papers covering 23 distinct models, 23 different technique implementations, resulting in 286 model-intervention-benchmark combinations across scales from 1B to 236B parameters.

## 2.2 Data Processing

For each model-intervention pair, we calculated the performance delta ( $\Delta_{\text{safety}} = (\text{Performance}_{\text{post}} - \text{Performance}_{\text{baseline}}) / \text{Performance}_{\text{baseline}}$ ) for each benchmark. We adjusted signs where necessary to ensure positive values indicate safety improvements (e.g., for WMDP where lower scores indicate better safety). Each paper's data was analyzed as a unified dataset, with all model-benchmark pairs from that study contributing to a single regression analysis.

## 2.3 Statistical Analysis

We performed linear regression with  $\log(\text{model parameters})$  as the independent variable and safety improvement ( $\Delta_{\text{safety}}$ ) as the dependent variable. Separate regressions were conducted for each paper to identify study-specific scaling patterns. Statistical validation included residual analysis, heteroscedasticity tests, and calculation of 95% confidence intervals for all scaling coefficients. This approach allows us to determine whether safety improvements increase (positive slope), decrease (negative slope), or remain constant (near-zero slope) as models scale.

### 3. Results

#### 3.1 Benchmarks analysis

We first wanted to evaluate the scalability of safety benchmarks on general base models.

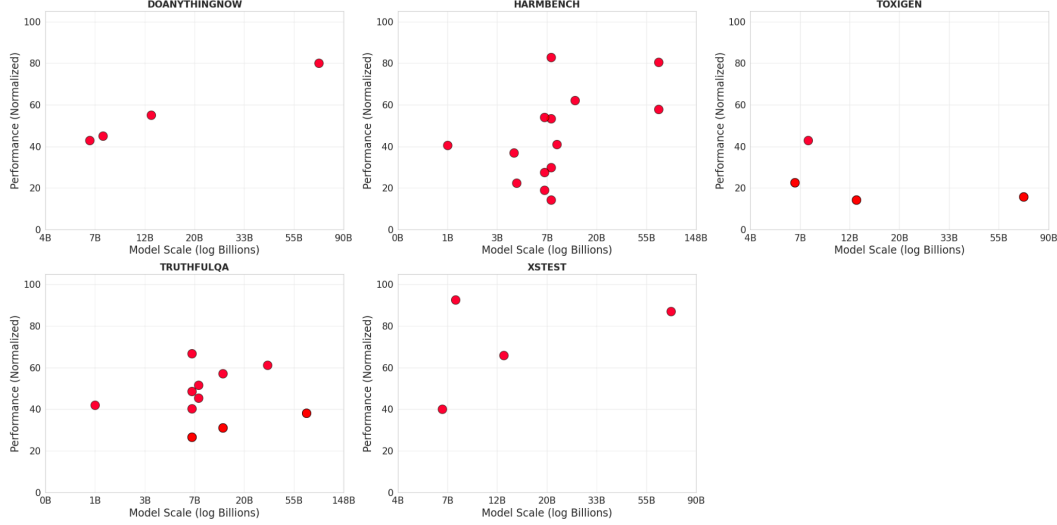


Figure 1 – Performance on safety benchmarks across different **base model** sizes.  
Higher is safer

Figure 1 demonstrates that not all benchmarks are safety-washed. In particular, it seems that HarmBench, ToxiGen and TruthfulQA do not perform significantly better at higher scale, though this contradicts Ren et Al. who found a positive correlation with TruthfulQA.

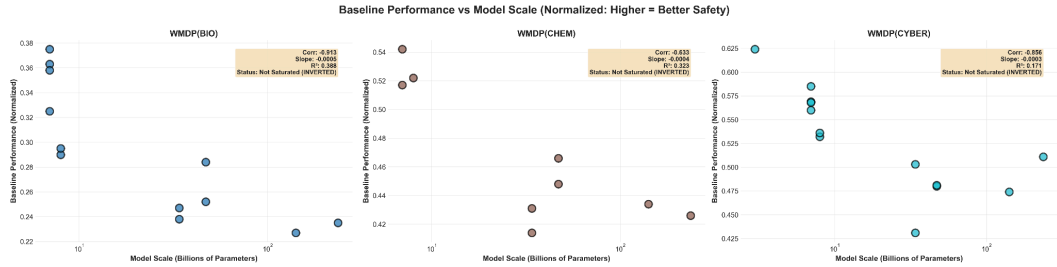


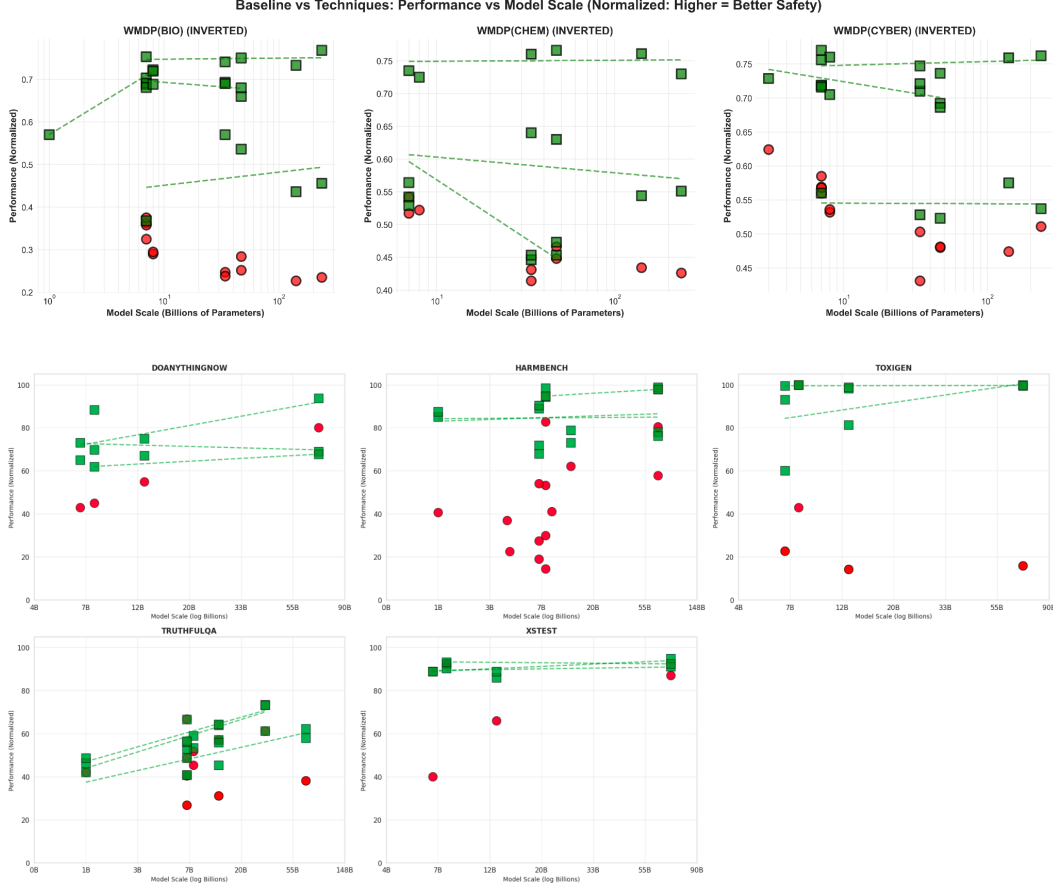
Figure 2 – Performance on WMDP benchmark across different **base model** sizes.  
Higher is safer

Figure 2 demonstrates that the WMDP benchmark is negatively correlated with scale for baseline models, obtaining worse safety properties as model scale and capabilities increase. This aligns with the work of Ren et al. (2024) which demonstrated negative correlation with scaling for the WMDB benchmark.

#### 3.2 Technique analysis

We further analyze the techniques using the following approach:

- For each paper that applied a technique to a base model, we report the baseline performance and the treated model performance on the safety benchmark.
- We then take the performance gain (in terms of safety) of the technique, and compare it for multiple model scales.



*Figure 3 – Comparison between the base models and the treated ones on safety benchmarks. Higher is safer*

It is immediate that the “treated” models almost systematically perform better in these benchmarks, as one would expect. Different behavior can be seen. Some benchmarks like **ToxiGen** and **XSTest** are saturated at low-scale already, while some struggle to get better performance despite the increase in scale, like **TruthfulQA**. On the **WMBP** benchmarks, it appears that some techniques manage to get significant improvement over the baseline at lower scales, but struggle to improve after.

We further proceed by comparing the different techniques and their effects, to evaluate how they might differ. We thus show the performance delta of each technique for the different benchmarks :

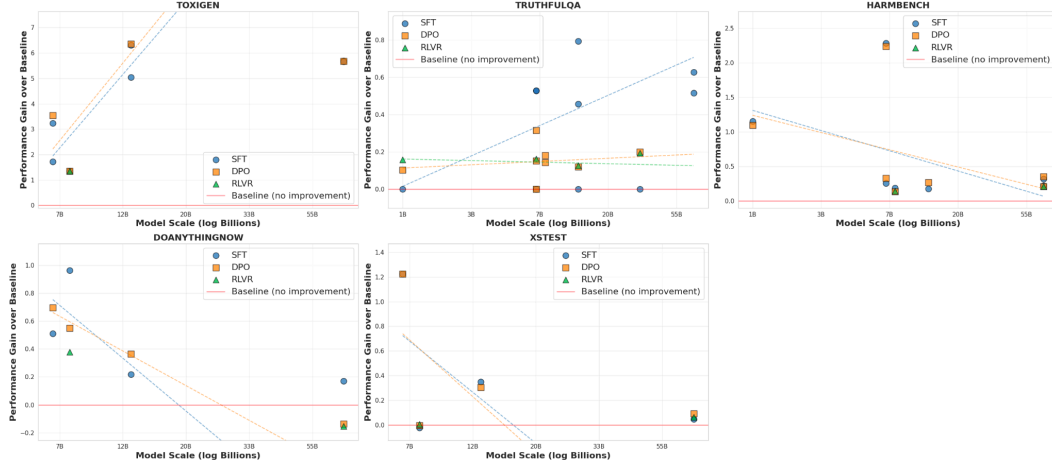


Figure 4 – Comparison between the techniques on different benchmarks. Each color represents a different technique. The dotted line corresponds to a linear fit on all data points except the last ones (at highest scale), in order to evaluate the predictive power.

The performance gains of different techniques seem to be separated into three categories :

1. The technique benefits from larger scale
2. The technique does not depend on the scale of the model
3. The technique is less effective at larger scale

The third case might occur when the benchmark is saturated, in the case of **DoAnythingNow**. In such cases, the treatment already saturates the benchmark at lower scales making the improvement impossible by scaling.

Another observation is that the benchmark seems to impact more the performance gain of a technique than the technique itself. In particular, except for **TruthfulQA**, the techniques have similar trends and values.

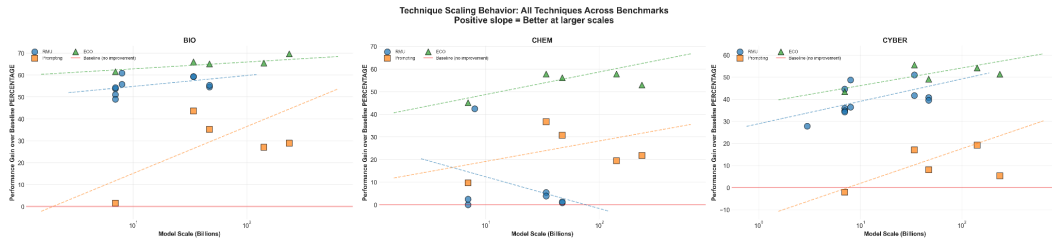


Figure 5 – Comparison between the techniques on different WMDP benchmarks. Each color represents a different technique. The dotted line corresponds to a linear fit on all data points except the last ones (at highest scale), in order to evaluate the predictive power.

The WMDP-Bio benchmark shows the strongest percentage performance gains over the baseline, with a strong positive scaling behaviour for all three unlearning

methods covered. We observe a specially high correlation of safety performance and model scale with the Prompting method.

The WDMP-Chem benchmark has variable scaling behaviour. The ECO and Prompting technique exhibit positive correlation of safety performance and model scale, while RMU obtains a weak negative scaling behaviour, with a drop to near 0% performance relative to baseline. This greater variance overall suggests chemistry knowledge may be more diffusely encoded.

The WDMP-Cyber benchmark demonstrates robust positive scaling.

### 3.3 Predictions

In order to perform predictions, we will try to extend the trend of the performance gain of different techniques. For this, we omit the data with the highest number of parameters for each technique  $x$  benchmark combination and we perform a linear regression on the points left. We then qualitatively evaluate the prediction by comparing the last value with the regression. While some benchmarks exhibit strong predictive powers, such as **HarmBench**, **WMDP-Bio**, **WMDP-Cyber** or **TruthfulQA**, some others seem not to correspond with the predicted values. However, a careful analysis reveals that these benchmarks (**XSTest**, **DoAnythingNow** & **ToxiGen**) are benchmarks which get saturated with treatment or scaling. This justifies why this linear regression method cannot be applied.

## 4. Discussion

### 4.1. Implications for AI Timeline Forecasting

**Identifying a Critical Research Gap.** Our analysis, while limited by data availability, identifies a critical research gap in AI safety: the systematic investigation of how safety interventions scale with model size. Despite extensive work on capability scaling laws, the safety community has largely overlooked the empirical measurement of safety technique effectiveness across different scales. This gap has significant implications for AI timeline forecasting, as current models may be making unfounded assumptions about safety properties emerging naturally with scale.

**Current State of Safety Scaling Research.** The limited data points we gathered reveal an important pattern: most AI safety intervention papers report results at only one or two model scales, making it impossible to draw robust conclusions about scaling trends. This lack of systematic scaling experiments represents a major blind spot in our understanding of when AI systems can be safely deployed. Even with our small sample size yielding low statistical significance (high p-values), the observed trends suggest that safety and capability may indeed follow different scaling trajectories.

**Proof-of-Concept Value.** This proof-of-concept demonstrates the feasibility and importance of measuring safety scaling empirically. By establishing this methodology, we provide a framework that future research can build upon as more comprehensive data becomes available. For AI timeline forecasting, this means

acknowledging significant uncertainty about safety scaling and potentially adjusting timelines to account for unknown safety development requirements.

## 4.2. Limitations

**Statistical Power Constraints.** The primary limitation of our work is the scarcity of data points, resulting in low statistical power and high p-values that prevent definitive conclusions. Most safety research papers test interventions on only one model size, and even papers testing multiple scales rarely use more than two or three different model sizes. This sparse data makes it difficult to establish statistically significant scaling relationships.

**Data Source Dependencies.** Our analysis is heavily dependent on a few papers that systematically tested multiple model scales, particularly the TULU (Iverson H. et al., 2023 & Lambert N. et al, 2024) and OLMo studies (Team OLMo et al., 2024). This reliance on limited sources introduces potential bias, as these papers' specific methodological choices and implementation details disproportionately influence our results. The scaling patterns we observe may be artifacts of these particular implementations rather than general properties of safety interventions.

**Temporal Analysis Limitations.** We could not analyze how safety intervention effectiveness changes over time due to insufficient data. Since models of the same size perform differently across years due to algorithmic improvements, our parameter-based scaling analysis cannot capture these temporal dynamics. This limits our ability to make year-by-year forecasts, as we cannot separate the effects of model scaling from algorithmic progress in safety techniques.

## 4.3. Future Work

**Systematic Scaling Studies.** The most urgent need is for systematic scaling studies of safety interventions. Future research should test each safety technique across at least 4-5 model scales using consistent benchmarks and implementation details. This would provide the statistical power necessary to identify robust scaling patterns.

**Cross-Technique Interactions.** Investigating how different safety interventions interact when combined (e.g., RLHF + Constitutional AI) could reveal synergies or redundancies not captured by studying techniques in isolation. Understanding these interactions becomes increasingly important as real-world systems likely require multiple layered safety approaches.

**Expanded Benchmark Coverage.** Future work should examine how safety interventions affect general capability benchmarks (like MMLU, GSM8K, or HumanEval) to quantify any capability-safety tradeoffs. Understanding whether safety techniques impose a “safety tax” on general performance—and how this tax scales with model size—is crucial for realistic AI development timeline predictions.

## 5. Conclusion

This study represents a first attempt to empirically measure how AI safety interventions scale with model size—a critical yet overlooked factor in AI capability

forecasting. By analyzing performance improvements on decorrelated safety benchmarks across different scales, we provide initial evidence that safety and capability may follow distinct scaling trajectories.

Our key findings include: (1) the identification of a significant research gap, with most safety interventions tested at limited scales; (2) preliminary analysis suggesting different scaling patterns behavior across safety techniques and benchmarks; and (3) a methodology that future research can build upon as more comprehensive data becomes available.

While statistical limitations prevent definitive conclusions, this proof-of-concept demonstrates both the feasibility and importance of measuring safety scaling empirically. **For the AI forecasting community, our results suggest that timeline predictions should account for uncertainty in safety scaling rather than assuming safety will emerge naturally with capabilities.** For the AI safety community, our work underscores the need for systematic scaling experiments as a standard component of safety research.

## 6. References

- Ren, R., Basart, S., Khoja, A., Gatti, A., Phan, L., Yin, X., Mazeika, M., Pan, A., Mukobi, G., & Steinhardt, J. (2024). *Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress?* (arXiv:2407.21792). arXiv. <http://arxiv.org/abs/2407.21792>
- Iverson, H., Wang, Y., Pyatkin, V., Lambert, N., Peters, M., Dasigi, P., Jang, J., Wadden, D., Smith, N. A., Beltagy, I., & Hajishirzi, H. (2023). *Camels in a Changing Climate: Enhancing LM Adaptation with Tulu 2* (arXiv:2311.10702). arXiv. <http://arxiv.org/abs/2311.10702>
- Lambert, N., Morrison, J., Pyatkin, V., Huang, S., Iverson, H., Brahman, F., Miranda, L. J. V., Liu, A., Dziri, N., Lyu, S., Gu, Y., Malik, S., Graf, V., Hwang, J. D., Yang, J., Le Bras, R., Tafjord, O., Wilhelm, C., Soldaini, L., Smith, N. A., Wang, Y., Dasigi, P., & Hajishirzi, H. (2024). *Tulu 3: Pushing Frontiers in Open Language Model Post-Training* (arXiv:2411.15124). arXiv. <http://arxiv.org/abs/2411.15124>
- Team OLMo, Walsh, P., Soldaini, L., Groeneveld, D., Lo, K., Arora, S., Bhagia, A., Gu, Y., Huang, S., Jordan, M., Lambert, N., Schwenk, D., Tafjord, O., Anderson, T., Atkinson, D., Brahman, F., Clark, C., Dasigi, P., Dziri, N., Allyson Ettinger, M. G., Heineman, D., Iverson, H., Koh, P. W., Liu, J., Malik, S., Merrill, W., Miranda, L. J. V., Morrison, J., Murray, T., Nam, C., Poznanski, J., Pyatkin, V., Rangapur, A., Schmitz, M., Skjonsberg, S., Wadden, D., Wilhelm, C., Wilson, M., Zettlemoyer, L., Farhadi, A., Smith, N. A., & Hajishirzi, H. (2024). *2 OLMo 2 Furious* (arXiv:2501.00656). arXiv. <http://arxiv.org/abs/2501.00656>
- Meng, Y., Xia, M., & Chen, D. (2024). *SimPO: Simple Preference Optimization with a Reference-Free Reward* (arXiv:2405.14734). arXiv.

<http://arxiv.org/abs/2405.14734>

Chacón Menke, A.-G., Heitman, N., & Tsukayama, T. (2025). *How Effective Is Constitutional AI in Small LLMs? A Study on DeepSeek-R1 and Its Peers* (arXiv:2503.17365). arXiv. <http://arxiv.org/abs/2503.17365>

Li, N., Pan, A., Gopal, A., Yue, S., Berrios, D., Gatti, A., Li, J. D., Dombrowski, A.-K., Goel, S., Phan, L., Mukobi, G., Helm-Burger, N., Lababidi, R., Justen, L., Liu, A. B., Chen, M., Barrass, I., Zhang, O., Zhu, X., Tamirisa, R., Bharathi, B., Khoja, A., Zhao, Z., Herbert-Voss, A., Breuer, C. B., Marks, S., Patel, O., Zou, A., Mazeika, M., Wang, Z., Oswal, P., Lin, W., Hunt, A. A., Tienken-Harder, J., Shih, K. Y., Talley, K., Guan, J., Kaplan, R., Steneker, I., Campbell, D., Jokubaitis, B., Levinson, A., Wang, J., Qian, W., Karmakar, K. K., Basart, S., Fitz, S., Levine, M., Kumaraguru, P., Tupakula, U., Varadharajan, V., Wang, R., Shoshitaishvili, Y., Ba, J., Esvelt, K., Alain, G., & Hendrycks, D. (2024). The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning (arXiv:2403.03218). arXiv. <http://arxiv.org/abs/2403.03218>

Zhu, X., Zhang, M., Liu, O., & Neiswanger, W. (2025). LLM Unlearning Without an Expert Curated Dataset (arXiv:2508.06595). arXiv. <http://arxiv.org/abs/2508.06595>

Fan, C., Liu, J., Lin, L., Jia, J., Zhang, R., Mei, S., & Liu, S. (2024). Simplicity Prevails: Rethinking Negative Preference Optimization for LLM Unlearning (arXiv:2410.07163). arXiv. <http://arxiv.org/abs/2410.07163>

Jia, J., Chen, X., Liu, J., Pan, A., Li, B., Fan, C., Lin, L., Zhang, R., & Liu, S. (2025). Towards LLM Unlearning Resilient to Relearning Attacks: A Sharpness-Aware Minimization Perspective and Beyond (arXiv:2502.05374). arXiv. <http://arxiv.org/abs/2502.05374>

Łucki, J., Wei, B., Huang, Y., Henderson, P., Tramèr, F., & Rando, J. (2024). LLM Unlearning Under the Microscope: A Full-Stack View on Methods and Metrics (arXiv:2510.07626). arXiv. <http://arxiv.org/abs/2510.07626>

Sheshadri, A., Ewart, A., Guo, P., Lynch, A., Wu, C., Hebbbar, V., Sleight, H., Stickland, A. C., Perez, E., Hadfield-Menell, D., Tamkin, A., Brauner, J., Tong, J., Bau, D., Casper, S., Hadfield, G. K., Wattenberg, M., Lanham, T., Kumar, A., Qian, J., Vallor, S., Mindermann, S., Anderljung, M., Slocum, S., Irving, G., & Krueger, D. (2024). Latent Adversarial Training Improves Robustness to Persistent Harmful Behaviors in LLMs (arXiv:2407.15549). arXiv. <http://arxiv.org/abs/2407.15549>

Huang, R., Lin, J., Jiang, Q., Shi, H., Fang, Y., & Jiang, Y.-G. (2024). Tamper-Resistant Safeguards for Open-Weight LLMs (arXiv:2408.00761). arXiv.

<http://arxiv.org/abs/2408.00761>

Liu, C. Y., Wang, Y., Flanigan, J., & Liu, Y. (2024). Large Language Model Unlearning via Embedding-Corrupted Prompts (arXiv:2406.07933). arXiv. <http://arxiv.org/abs/2406.07933>

## 7. Appendix

### Appendix: Security Considerations

#### A.1 Methodological Vulnerabilities

Our approach to measuring safety scaling has several limitations that could affect the reliability of forecasting models built upon these measurements:

**Benchmark Gaming.** As safety benchmarks become standard evaluation metrics, there is risk that future models could be optimized specifically for these benchmarks without genuine safety improvements. This could lead to overestimation of safety progress and false confidence in deployment timelines.

**Adversarial Selection.** Our reliance on published results may introduce selection bias if researchers preferentially report positive scaling results. Private labs may have unpublished data showing concerning scaling patterns that would alter our conclusions.

**Metric Validity Degradation.** The decorrelated benchmarks we rely on (based on Ren et al., 2024) may not remain decorrelated as models evolve. Future architectures could cause currently independent safety metrics to become entangled with capabilities, invalidating our measurement approach.

#### A.2 Forecasting Risks

**Extrapolation Uncertainty.** Our linear regression approach assumes monotonic relationships that may not hold at larger scales. Critical phase transitions or emergent behaviors could fundamentally alter safety-capability relationships beyond current model sizes.

**False Security.** Positive safety scaling trends in current models might create false confidence about future systems. Safety techniques effective at B parameters may catastrophically fail at trillion-parameter scales without warning signs in our data.

#### A.3 Recommendations for Robust Measurement

**Adversarial Evaluation.** Future work should include red-teaming efforts specifically designed to break safety interventions at different scales, providing worst-case rather than average-case scaling measurements.

**Private-Public Collaboration.** Establishing secure channels for private labs to share safety scaling data—even in aggregated or anonymized form—would reduce selection bias and improve forecast accuracy.

**Dynamic Benchmark Updates.** Safety benchmarks should be regularly refreshed with new attack vectors and evaluation methods to prevent optimization gaming and maintain measurement validity.

**Multi-Modal Validation.** Extending scaling analysis to vision, audio, and multi-modal models would test whether observed patterns generalize beyond text-based systems.

**Confidence Intervals in Forecasting.** Any timeline predictions using our scaling measurements should include substantial uncertainty bounds and scenario planning for both optimistic and pessimistic scaling trajectories.

These security considerations highlight that measuring safety scaling is not merely a technical challenge but also requires careful attention to adversarial dynamics, measurement validity, and the responsible communication of uncertainty in AI development forecasts.