Foundational Models Must Be Designed To Yield Safer Loss Landscapes That Resist Harmful Fine-Tuning

Karan Uppal
karan.uppal3@gmail.com

Pavan Kalyan
tankalapavankalyan@gmail.com

Abstract

This position paper argues that foundational models must be engineered during pretraining to develop inherent resistance to harmful fine-tuning, rather than relying on post-training interventions or inference-time guardrails. Recent works have shown that even minimal adversarial data can readily compromise safety alignment in state-of-the-art models at remarkably low cost. We propose an integrated approach combining loss landscape engineering, self-destructing model techniques, and constrained optimization to create models that naturally resist harmful adaptations while preserving beneficial fine-tuning capabilities. By proactively addressing this vulnerability through pretraining interventions rather than reactive measures, we can enhance the safety and trustworthiness of AI systems as they continue to advance in capabilities.

1. Introduction

Foundational models are advancing rapidly, showing exceptional capabilities in multiple fields. However, their susceptibility to harm fine-tuning poses a major security risk.

> **Foundational models must be engineered during pretraining to develop inherent resistance to harmful fine-tuning, rather than relying on post-training interventions or inference-time guardrails.**

The range of induced harmful behaviors is broad, including malware generation, hate speech, and instructions for dangerous activities (AI Safety Group, 2025). As model capabilities improve with scale, they can become more difficult to defend and more powerful when operated with malicious intent (Halawi et al., 2024). Recent works (Qi et al., 2023; Zhan et al., 2023) have shown the ease with which safety mechanisms in foundational models can be compromised through fine-tuning. This vulnerability exists for both open-source models and closed-access models that offer fine-tuning APIs. Even fine-tuning with seemingly benign datasets, intended to improve a model's responsiveness, can inadvertently degrade safety guardrails (Qi et al., 2023). The current safety paradigm fails to address the vulnerability at a fundamental level. This creates a tension between model adaptability and safety that cannot be resolved through post-training measures alone (Halawi et al., 2024).

Current safety mechanisms mainly focus on integrating safety rules within pretrained models to constrain model behavior during inference (Bai et al., 2022b; Wang et al., 2024). These inference-time guardrails, although beneficial, can be easily bypassed by fine-tuning. Additionally, moderation systems designed to filter out harmful content during fine-tuning can be bypassed using adversarial techniques (Wallace et al., 2019). The problem lies in the reactive nature of current approaches, that treat safety as an add-on property rather than an inherent characteristic of the model. By shaping the loss landscape during pretraining, we can build models that resist harmful adaptations by design, while remaining open to beneficial fine-tuning, addressing vulnerabilities at their root rather than applying post hoc fixes. In this position paper, we argue for pretraining-time interventions as the most effective path to safety. We critique current approaches, present an integrated framework combining loss landscape engineering, self-destructing models, and constrained optimization, and address potential criticisms, outlining key directions for future research in building inherently safe foundation models.

2. Safety: Pretraining vs Post-training

We begin by presenting our views on first alternative perspective to our position: Safety through post-training interventions. Section 2.1 begins by examining and critiquing the current paradigm of post-training interventions for AI safety. Section 2.2 presents initial research that demonstrates improved safety through pretraining techniques.

2.1. Existing Safety Methods and Limitations

The AI safety community has explored various strategies to mitigate risks from foundation models, including inference-time guardrails like constitutional AI (Bai et al., 2022b), data filtering, and safety-specific continued pretraining. While these methods can block obvious misuse, they mostly operate at the output level and are vulnerable to fine-tuning attacks. Qi et al. (2023) has shown that just 15 targeted examples can override safety guardrails. Content filters may miss subtle or novel forms of harmful content. Adversarial techniques can generate training examples that appear benign to filtering systems while still inducing harmful behaviors when used for fine-tuning (Huang et al., 2025). Post-training safety measures like continued pretraining on safety data and RLHF (Bai et al., 2022a) aim to align models with human values but often trade off adaptability for safety. Despite their complexity, these methods remain vulnerable to fine-tuning attacks (Zhan et al., 2023), failing to address the core issue: the model's parameter space remains open to harmful adaptation.

As noted in Marcus et al. (2020), "current filtering approaches operate on a blacklist principle, making them inherently reactive and unable to anticipate novel forms of harmful content". The core issue is that post-training safety measures operate on top of a foundation that is still adaptable. They attempt to constrain model outputs without addressing the underlying sensitivity of the parameter space to harmful adaptation. As AI systems become more powerful and widely deployed, this vulnerability becomes increasingly critical to address at a fundamental level (Halawi et al., 2024).This represents a limitation of current safety approaches that can only be overcome through interventions during the pretraining phase (Cao et al., 2023).

2.2. Early Evidence for Pretraining Interventions

Recent works like self-destructing models (Henderson et al., 2023) address safety during the pretraining phase by creating inherent resistance to certain adaptations while preserving beneficial capabilities. They are designed to function normally for intended applications while strategically "self-destructing" when attempts are made to adapt them for harmful purposes (Henderson et al., 2023). Moreover, Peng et al. (2024) has identified "safety basins" - parameter regions where random perturbations maintain alignment, demonstrating inherent safety preservation mechanisms in pretrained models. This is supported by recent literature suggesting that, analyzing loss landscapes and optimization trajectories can provide insights into why pretraining-then-fine-tuning paradigms improve performance and generalization capability across different tasks (Hao et al., 2019). Mehta et al. (2021) has shown that pretrained weights appear to ease forgetting by leading to wider minima in the loss landscape, suggesting that careful pretraining can create more robust model. Recent work on pretraining poisoning offers additional evidence on the importance of the pretraining phase for model safety. Zhang et al. (2024) has shown that even minimal poisoning during pretraining ($< 0.1\%$ of the data) can persist after models are fine-tuned as helpful and harmless chatbots. These findings collectively indicate that incorporating safety measures during pretraining lead to a more fundamental safety understanding as compared to post-training strategies. By shaping the very terrain of the parameter space, pretraining-based approaches can create models that naturally resist harmful adaptations while remaining receptive to beneficial fine-tuning.

3. Proposal

We propose a proactive approach to model safety by shaping the loss landscape during pretraining. The aim is to incorporate safety properties directly into the model so that they persist even after fine-tuning. We build on ideas from loss landscape geometry, adversarial training, meta-learning, and constrained optimization to propose a framework for training conditionally tamper-resistant models.

3.1. Motivation: Self-Destructing Models

Our proposal is motivated by the concept of self-destructing models, which are designed to resist adaptation for harmful purposes while remaining effective for beneficial applications. The core insight is that by engineering the loss landscape during pretraining, we can create models that are conditionally tamper-resistant. This approach addresses several limitations of current safety methods:

1. **Persistence under fine-tuning**: Unlike inference-time guardrails that can be easily bypassed through fine-tuning, tamper-resistant models preserve their safety properties when their parameters are modified.
2. **Proactive rather than reactive**: Instead of filtering out harmful content during inference, this approach proactively prevents harmful model adaptation.
3. **Reduced arms race dynamics**: Integrating resistance within the model reduces the arms race between safety mechanisms and techniques to circumvent them.
4. **Compatibility with open-source deployment**: Resulting models can be safely released as open-source, as their resistance to harmful adaptation is inherent.

Instead of creating models that resist all forms of adaptation, which would limit their usefulness, we propose creating models that selectively resist harmful adaptations while remaining receptive to beneficial ones. This selective resistance can be achieved by engineering the loss landscape during pretraining, creating a foundation for AI systems that are inherently safer by design while maintaining their flexibility for beneficial applications.

3.2. Core Idea: Conditional Tamper Resistance

We propose a framework for creating conditionally tamper-resistant pretrained models by engineering the loss landscape. This framework integrates several complementary approaches to work in unison to build models with conditional tamper resistance.

Loss landscape engineering: We propose creating high-energy barriers around harmful regions of the parameter space, making them difficult to reach through standard optimization processes, resulting in increased difficulty for harmful fine-tuning. Sharpness-aware minimization principle (Foret et al., 2020) is one example, where loss landscape is engineered to improve generalization.

Self-destructing model techniques: Building on the work of Henderson et al. (2023), we propose incorporating self-destructing mechanisms that activate when attempts are made to adapt the model for harmful purposes. These mechanisms leverage meta-learning and adversarial training to create models that resist specific types of adaptations.

Constrained optimization: By incorporating safety constraints directly into the optimization process during pretraining, we can create models that naturally gravitate toward safe regions of the parameter space.

The key innovation in our proposal is the integration of these approaches to create models that are conditionally tamper-resistant. In contrast to earlier work, which focuses on either post-training safety measures or complete resistance to adaptation, our approach proposes to create models that selectively resist harmful adaptations while remaining receptive to beneficial ones. Huang et al. (2024) has demonstrated progress in this direction, showing that attenuating harmful perturbation over model weights can help create models that resist harmful fine-tuning while maintaining performance on beneficial downstream tasks. The resulting models would be safe to release as open-source, as their resistance to harmful adaptation is integrated within their parameters. Our proposal addresses a critical challenge in AI safety: how to balance the benefits of open-source development with the risks of misuse.

3.3. Approaches and Foundations

i) Existing methods Several existing methods provide foundations for creating conditionally tamper-resistant models through loss landscape engineering. **Meta-Learned Adversarial Censoring (MLAC)**, introduced by Henderson et al. (2023), using both meta-learning and adversarial learning, demonstrated that MLAC could prevent a BERT-style model from being repurposed for gender identification (the harmful task in this context) while preserving its ability to perform profession classification (the desirable task). **Tamper-Resistant Safeguards (TAR)** creates models that actively resist modification of their safety properties. TAR incorporates adversarial training and specialized loss functions to create high-energy barriers around harmful regions of the parameter space (Tamirisa et al., 2024). **Safety pretraining** as proposed by Maini et al. (2025) has demonstrated that interventions during pretraining can create models with inherent resistance to harmful fine-tuning across a wide range of potential attack scenarios.

ii) Connection to loss landscapes The connection between these approaches and loss landscape engineering is significant. Loss landscapes in neural networks can be visualized as high-dimensional terrains with valleys (local minima), ridges, and plateaus (Li et al., 2018). The shape of this terrain determines how easily a model can be adapted for different purposes through fine-tuning. When models are fine-tuned from pretrained weights, they tend to stay in the same basin in the loss landscape and stay close in parameter space (Neyshabur et al., 2020). Self-destructing models can exploit this property by creating basins that are difficult to escape when attempting certain adaptations but conducive to desired tasks. High-energy barriers represent regions of higher loss that separate different basins in the loss landscape. By creating such barriers around harmful regions of the parameter space, we can make it difficult for fine-tuning to adapt the model for harmful purposes. Tamirisa et al. (2024) has demonstrated that carefully designed loss landscape barriers can create models with selective resistance to harmful adaptations while maintaining full receptivity to beneficial fine-tuning. Their work provides empirical evidence for the effectiveness of this approach, showing that models with engineered loss landscapes can maintain safety properties even after extensive fine-tuning attempts specifically designed to compromise them.

3.4. Open Research Questions

We identify five research directions that define a comprehensive agenda for safe model deployment. Progress on these fronts will not only clarify what is possible in tamper resistance but also provide the theoretical grounding needed to design, evaluate, and trust AI systems.

1. **Quantifying Minimal Units of Tamper Resistance**
   a) What is the smallest parameter-space perturbation that reliably blocks harmful adaptation?
   b) How do geometric invariants (e.g., curvature, connectivity) of loss basins relate to empirical measures of resistance?
2. **Identifying Geometric Signatures of Unsafe Adaptation**
   a) Are there universal geometric features that distinguish safe adaptation paths from unsafe ones?

b) Can metastable loss landscape regions serve as indicators of vulnerability to harmful fine-tuning?

3. **Understanding Temporal Causality in Landscape Interventions**
   a) Is there a critical window during training when interventions have the most impact on safety?
   b) Does early intervention (e.g., initialization) create more stable resistance than late-stage regularization?
4. **Forming Scaling Laws for Landscape Engineering**
   a) How does tamper resistance scale with model size, depth, or capacity?
   b) What role does the intrinsic dimensionality of solution manifolds play in constraining landscape shaping techniques?
5. **Establishing Theoretical Limits**
   a) Can we use the complexity of harmful tasks to estimate difficulty to build resistance into the model?
   b) Is there a principle balancing precision in task preservation vs. harmful adaptation blocking?

These questions aim to advance loss landscape engineering from preliminary studies to scientific discipline, laying the theoretical groundwork for tamper-resistant AI systems.

4. Alternate views

We contrast our proposal with alternative perspectives and critiques, as well as respond to these concerns. Section 2.1 presents a detailed overview of why post-training alignment is not sufficient. Here, we present other alternatives along with our responses to their limitations.

4.1. Runtime Monitoring and Intervention

Criticism: Runtime monitoring detects and blocks harmful model use by inspecting inputs and outputs, acting independently of the model.

Response: While valuable, such systems are reactive, can be bypassed by adversarial inputs, and introduce latency. They also don't mitigate risks from harmful fine-tuning. A robust safety approach combines runtime monitoring with inherently safer models.

4.2. Access Controls and Deployment Restrictions

Criticism: Restricting model access and deployment can prevent misuse by limiting who can fine-tune or operate powerful models.

Response: Such restrictions trade off safety for openness. As compute becomes more accessible and open-source models proliferate, safety must be ensured at the model level. Incorporating safeguards during pretraining enables open release without compromising safety. Tamirisa et al. (2024) offers a technical solution that can preserve safety even with full weight access.

4.3. Computational Cost and Complexity

Criticism: Engineering loss landscapes during pretraining increases computational demands.

Response: Though more costly upfront, safer pretraining reduces long-term risks and the ongoing burden of post-deployment safety measures. As methods mature, efficiency will improve, broadening access. Once pretrained, these models can serve as safe, reusable foundations—lowering the barrier for smaller organizations without sacrificing safety.

4.4. Potential Reduction in Model Capabilities

Criticism: Constraining models via loss landscape engineering may limit their ability to adapt to useful, unforeseen tasks.

Response: Our goal is selective, not absolute resistance, i.e. targeting harmful adaptations while preserving beneficial flexibility. Empirical results (Henderson et al., 2023; Tamirisa et al., 2024; Maini et al., 2025) show this balance is achievable with minimal trade-offs. As techniques improve, we expect even finer control over adaptation, ensuring safety without sacrificing utility.

5. Conclusion

We argue that foundational models should be made inherently resistant to harmful fine-tuning by engineering safety during pretraining, rather than relying on post-training fixes or run-time guardrails. Current approaches leave models vulnerable due to their flexibility to adapt. By shaping the loss landscape during pretraining, we can create models that resist malicious adaptations while retaining beneficial ones. Our framework combines loss landscape engineering, self-destructing models, and constrained optimization to build conditionally tamper-resistant systems. We outline key open research questions to advance this direction. Although we recognize the criticisms, we believe that pretraining-based safety offers the most robust defense against misuse. By integrating safety measures within the foundations of AI systems, we can ensure that these powerful systems benefit humans while minimizing potential harms. We strongly urge the research community to invest in this proactive approach.

References

AI Safety Group. International ai safety report 2025. Technical report, International AI Safety Consortium, 2025.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova Dassarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, John Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862, 2022a. URL https://api.semanticscholar.org/CorpusID:248118878.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022b. URL https://arxiv.org/abs/2212.08073.

Yuanpu Cao, Bochuan Cao, and Jinghui Chen. Stealthy and persistent unalignment on large language models via backdoor injections. In *North American Chapter of the Association for Computational Linguistics*, 2023. URL https://api.semanticscholar.org/CorpusID:265551434.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Danny Halawi, Alexander Wei, Eric Wallace, Tony T. Wang, Nika Haghtalab, and Jacob Steinhardt. Covert malicious finetuning: Challenges in safeguarding llm adaptation, 2024. URL https://arxiv.org/abs/2406.20053.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Visualizing and understanding the effectiveness of bert. *ArXiv*, abs/1908.05620, 2019. URL https://api.semanticscholar.org/CorpusID:199668646.

Peter Henderson, Eric Mitchell, Christopher Manning, Dan Jurafsky, and Chelsea Finn. Self-destructing models: Increasing the costs of harmful dual uses of foundation models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 287–296, 2023.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Booster: Tackling harmful fine-tuning for large language models via attenuating harmful perturbation. *arXiv preprint arXiv:2409.01586*, 2024.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Virus: Harmful fine-tuning attack for large language models bypassing guardrail moderation. *ArXiv*, abs/2501.17433, 2025. URL https://api.semanticscholar.org/CorpusID:275954444.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

Pratyush Maini, Sachin Goyal, Dylan Sam, Alexander Robey, Yash Savani, Yiding Jiang, Andy Zou, Zacharcy C. Lipton, and J. Zico Kolter. Safety pretraining: Toward the next generation of safe ai. 2025. URL https://api.semanticscholar.org/CorpusID:278032920.

J. Scott Marcus, Georg Zachmann, Stephen Gardner, Simone Tagliapietra, and Elissavet Lykogianni. Promoting product longevity. Study PE 648.767, Policy Department for Economic, Scientific and Quality of Life Policies, European Parliament, 2020. URL https://www.europarl.europa.eu/RegData/etudes/STUD/2020/648767/IPOL_STU(2020)648767_EN.pdf. Requested by the IMCO Committee.

Sanket Vaibhav Mehta, Darshan Patil, Sarath Chandar, and Emma Strubell. An empirical investigation of the role of pre-training in lifelong learning. *J. Mach. Learn. Res.*, 24:214:1–214:50, 2021. URL https://api.semanticscholar.org/CorpusID:245329773.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? *ArXiv*, abs/2008.11687, 2020. URL https://api.semanticscholar.org/CorpusID:221319407.

Sheng Y Peng, Pin-Yu Chen, Matthew Hull, and Duen H Chau. Navigating the safety landscape: Measuring risks in finetuning large language models. *Advances in Neural Information Processing Systems*, 37:95692–95715, 2024.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *ArXiv*, abs/2310.03693, 2023. URL https://api.semanticscholar.org/CorpusID:263671523.

Rishub Tamirisa, Bhrugu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, Justin Wang, Rowan Wang, Ron Arel, et al. Tamper-resistant safeguards for open-weight llms. *arXiv preprint arXiv:2408.00761*, 2024.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing nlp. *arXiv preprint arXiv:1908.07125*, 2019.

Jiongxiao Wang, Jiazhao Li, Yiquan Li, Xiangyu Qi, Junjie Hu, Yixuan Li, Patrick McDaniel, Muhao Chen, Bo Li, and Chaowei Xiao. Mitigating fine-tuning based jailbreak attack with backdoor enhanced safety alignment. *arXiv preprint arXiv:2402.14968*, 2024.

Qiusi Zhan, Richard Fang, Rohan Bindu, Akul Gupta, Tatsunori Hashimoto, and Daniel Kang. Removing rlhf protections in gpt-4 via fine-tuning. In *North American Chapter of the Association for Computational Linguistics*, 2023. URL https://api.semanticscholar.org/CorpusID:265067269.

Yiming Zhang, Javier Rando, Ivan Evtimov, Jianfeng Chi, Eric Michael Smith, Nicholas Carlini, Florian Tramèr, and Daphne Ippolito. Persistent pre-training poisoning of llms. *ArXiv*, abs/2410.13722, 2024. URL https://api.semanticscholar.org/CorpusID:273403633.