Novelty Assessment Report

Paper: Self-Destructive Language Models
PDF URL: https://openreview.net/pdf?id=ERNpUGr8M5
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

Harmful fine-tuning attacks represent a major threat to the security of large language models (LLMs), allowing adversaries to compromise safety guardrails with minimal harmful data. While existing defenses attempt to reinforce LLM alignment, they fail to address models' inherent `trainability' on harmful data, leaving them vulnerable to stronger attacks with increased learning rates or larger harmful datasets. To overcome this limitation, we introduce SEAM, a novel alignment-enhancing defense that transforms LLMs into self-destructive models with intrinsic resilience to misalignment attempts. Specifically, these models retain their capabilities for legitimate tasks while exhibiting substantial performance degradation when fine-tuned on harmful data. The protection is achieved through a novel loss function that couples the optimization trajectories of benign and harmful data, enhanced with adversarial gradient ascent to amplify the self-destructive effect. To enable practical training, we develop an efficient Hessian-free gradient estimate with theoretical error bounds. Extensive evaluation across LLMs and datasets demonstrates that SEAM creates a no-win situation for adversaries: the self-destructive models achieve state-of-the-art robustness against low-intensity attacks and undergo catastrophic performance collapse under high-intensity attacks, rendering them effectively unusable. The code is available: https://anonymous.4open.science/r/seam-5C7E (warning: this paper contains potentially harmful content generated by LLMs.)

Core Task Landscape

This paper addresses: **Defending Large Language Models Against Harmful Fine-Tuning Attacks**

A total of **50 papers** were analyzed and organized into a taxonomy with **20 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Alignment-Stage Defense Mechanisms**
- **Runtime Defense Mechanisms**
- **Post-Fine-Tuning Recovery and Mitigation**
- **Adversarial Training and Robustness Enhancement**
- **Attack Characterization and Threat Analysis**
- **Comprehensive Security Frameworks and Surveys**
- **Red Teaming and Attack Generation**

Complete Taxonomy Tree

- Defending Large Language Models Against Harmful Fine-Tuning Attacks Survey Taxonomy
- Alignment-Stage Defense Mechanisms
  - Perturbation-Based Alignment Enhancement ★ (5 papers)
  - [0] Self-Destructive Language Models (Anon et al., 2026) View paper
  - [1] Targeted vaccine: Safety alignment for large language models against harmful fine-tuning via layer-wise perturbation (Liu Guozhi, 2025) View paper
  - [5] Immunization against harmful fine-tuning attacks (Rosati, 2024) View paper
  - [12] Representation noising: A defence mechanism against harmful finetuning (David Atanasov, 2024) View paper
  - [33] Vaccine: Perturbation-aware Alignment for Large Language Models against Harmful Fine-tuning Attack (Sihao Hu, 2024) View paper
  - Gradient-Based Alignment Optimization (4 papers)
  - [10] Sdd: Self-degraded defense against malicious fine-tuning (Zixuan Chen, 2025) View paper
  - [20] Lisa: Lazy Safety Alignment for Large Language Models against Harmful Fine-tuning Attack (Sihao Hu, 2024) View paper
  - [24] Gradient surgery for safe llm fine-tuning (Yi Biao, 2025) View paper
  - [35] Lazy Safety Alignment for Large Language Models against Harmful Fine-tuning (Huang Tiansheng, 2024) View paper
  - Safety Data Curation and Augmentation (3 papers)
  - [16] Adaptive defense against harmful fine-tuning for large language models via bayesian data scheduler (Hu, 2025) View paper
  - [21] Safety fine-tuning at (almost) no cost: A baseline for vision large language models (Zong, 2024) View paper
  - [41] Pharmacist: Safety Alignment Data Curation for Large Language Models against Harmful Fine-tuning (Liu Guo-zhi, 2025) View paper
  - Tamper-Resistant Safeguard Integration (2 papers)
  - [19] On weaponization-resistant large language models with prospect theoretic alignment (Z Cheng, 2025) View paper
  - [49] Tamper-resistant safeguards for open-weight llms (Tamirisa, 2024) View paper
- Runtime Defense Mechanisms
  - Inference-Time Detection and Filtering (4 papers)
  - [2] Safeguarding large language models in real-time with tunable safety-performance trade-offs (Fonseca, 2025) View paper

    - [11] Scam Shield: Multi-Model Voting and Fine-Tuned LLMs Against Adversarial Attacks (Chang Chen-Wei, 2025) View paper
    - [15] Helping Large Language Models Protect Themselves: An Enhanced Filtering and Summarization System (Muhaimin, 2025) View paper
    - [26] Robust safety classifier for large language models: Adversarial prompt shield (Kim Jin-Hwa, 2023) View paper
    - Input Perturbation Defenses (2 papers)
    - [34] Large language model sentinel: Llm agent for adversarial purification (Lin Guang, 2024) View paper
    - [43] SmoothLLM: Defending Large Language Models Against Jailbreaking Attacks (Robey, 2023) View paper
    - Self-Evaluation Mechanisms (1 papers)
    - [47] Self-evaluation as a defense against adversarial attacks on llms (Brown, 2024) View paper
- Post-Fine-Tuning Recovery and Mitigation
    - Backdoor Detection and Unalignment Defense (4 papers)
    - [28] Probe before you talk: Towards black-box defense against backdoor unalignment for large language models (Yi Biao, 2025) View paper
    - [30] MetaDefense: Defending Finetuning-based Jailbreak Attack Before and During Generation (Jiang, 2025) View paper
    - [40] Backdooralign: Mitigating fine-tuning based jailbreak attack with backdoor enhanced safety alignment (Muhao Chen, 2024) View paper
    - [45] Mitigating fine-tuning based jailbreak attack with backdoor enhanced safety alignment (Wang, 2024) View paper
    - Prompt Template Preservation (1 papers)
    - [42] Keeping llms aligned after fine-tuning: The crucial role of prompt templates (Lyu Kaifeng, 2024) View paper
    - Safety Restoration Strategies (2 papers)
    - [18] Safe lora: The silver lining of reducing safety risks when finetuning large language models (Hsu, 2024) View paper
    - [46] Mitigating fine-tuning risks in llms via safety-aware probing optimization (Zhang Zhixin, 2025) View paper
- Adversarial Training and Robustness Enhancement
    - Adversarial Fine-Tuning for Jailbreak Defense (3 papers)
    - [9] Defending against alignment-breaking attacks via robustly aligned llm (Cao, 2024) View paper
    - [23] Baseline defenses for adversarial attacks against aligned language models (Jain, 2023) View paper
    - [37] Adversarial tuning: Defending against jailbreak attacks for llms (Liu Fan, 2024) View paper
    - Multimodal Adversarial Training (2 papers)
    - [17] Securing vision-language models with a robust encoder against jailbreak and adversarial attacks (Md. Zarif Hossain, 2024) View paper
    - [29] Adversarial training for multimodal large language models against jailbreak attacks (Lu LiMing, 2025) View paper
    - Pre-Trained Model Adversarial Adaptation (1 papers)
    - [31] How should pre-trained language models be fine-tuned towards adversarial robustness? (Tuan, 2021) View paper
- Attack Characterization and Threat Analysis
    - Fine-Tuning Attack Mechanisms and Vulnerabilities (4 papers)
    - [4] Fine-tuning aligned language models compromises safety, even when users do not intend to! (Qi, 2023) View paper
    - [14] Towards understanding the fragility of multilingual llms against fine-tuning attacks (Poppi, 2025) View paper
    - [38] Harmful fine-tuning attacks and defenses for large language models: A survey (Huang Tiansheng, 2024) View paper
    - [39] Analysing Safety Risks in LLMs Fine-Tuned with Pseudo-Malicious Cyber Security Data (Arief, 2025) View paper
    - Jailbreak Attack Taxonomy and Analysis (3 papers)
    - [6] Jailbreak attacks and defenses against large language models: A survey (Sibo Yi, 2024) View paper
    - [13] Adversarial attacks and defenses for large language models (LLMs): methods, frameworks & challenges (Pranjal Kumar, 2024) View paper
    - [44] Attack prompt generation for red teaming and defending large language models (Deng Boyi, 2023) View paper
    - Domain-Specific Threat Assessment (2 papers)
    - [3] Open-weight genome language model safeguards: Assessing robustness via adversarial fine-tuning (James R. M. Black, 2025) View paper
    - [25] Unified Defense for Large Language Models against Jailbreak and Fine-Tuning Attacks in Education (Xin Yi, 2025) View paper
- Comprehensive Security Frameworks and Surveys
    - Backdoor Threats and Defenses (1 papers)
    - [8] A survey on backdoor threats in large language models (llms): Attacks, defenses, and evaluations (Zhou Yihe, 2025) View paper
    - General LLM Security Surveys (2 papers)
    - [7] Securing large language models: Threats, vulnerabilities and responsible practices (Abdali, 2024) View paper
    - [36] Safety at scale: A comprehensive survey of large model safety (Ma, 2025) View paper
    - Dual-Use Risk Assessment (2 papers)
    - [27] Dual-Use of Large Language Models (LLMs) and Generative AI (GenAI) in Cybersecurity: Risks, Defenses, and Governance Strategies (Kiarash Ahi, 2025) View paper
    - [48] Large Language Models (LLMs) and Generative AI in Cybersecurity and Privacy: A Survey of Dual-Use Risks, AI-Generated Malware, Explainability, and Defensive â (K Ahi, 2025) View paper
- Red Teaming and Attack Generation (3 papers)
    - [22] Threat Landscape of Adversarial Attacks on Generative AI and Large Language Models (LLMs): Exploring Different Types of Adversarial Attacks, Associated Risks â (I Naik, 2025) View paper
    - [32] Adversarial Attacks on Large Language Modelâ□□Based System and Mitigating Strategies: A Case Study on ChatGPT (Bowen Liu, 2023) View paper
    - [50] Ethical Treatment of Language Models Against Harmful Inference-Time Interventions (JesÃºs Fernando Cevallos Moreno, 2025) View paper

Narrative

Core task: Defending large language models against harmful fine-tuning attacks. The field has organized itself around several complementary strategies for protecting aligned LLMs from adversarial fine-tuning that could restore harmful capabilities. At the highest level, the taxonomy distinguishes between defenses applied during alignment (Alignment-Stage Defense Mechanisms), those that operate at inference time (Runtime Defense Mechanisms), and methods for recovering models after compromise (Post-Fine-Tuning Recovery and Mitigation). Additional branches address proactive robustness building (Adversarial Training and Robustness Enhancement), understanding attack surfaces (Attack Characterization and Threat Analysis), holistic protection schemes (Comprehensive Security

Frameworks and Surveys), and offensive security research (Red Teaming and Attack Generation). Within alignment-stage defenses, researchers have explored diverse approaches including perturbation-based methods that inject noise or modify representations during training, as exemplified by works like Immunization against harmful fine-tuning[5] and Representation noising[12], alongside vaccine-style interventions such as Targeted vaccine[1] and Vaccine[33] that preemptively inoculate models against specific attack patterns.

A particularly active line of inquiry focuses on making safety alignment more robust to fine-tuning degradation, with many studies examining trade-offs between model utility and resistance to adversarial updates. Self-Destructive Language Models[0] sits within the perturbation-based alignment enhancement cluster, sharing conceptual ground with Immunization against harmful fine-tuning[5] and Representation noising[12], all of which modify internal model states or training dynamics to preserve safety properties. These approaches contrast with vaccine methods like Targeted vaccine[1], which tend to emphasize curated adversarial examples rather than architectural or representational interventions. Meanwhile, runtime defenses such as Jailbreak attacks and defenses[6] and post-hoc recovery techniques address orthogonal threat windows, and comprehensive surveys like Safeguarding large language models[2] attempt to synthesize insights across these diverse protection paradigms. Open questions remain around scalability, the balance between safety and capability retention, and whether any single defense layer suffices against sophisticated adaptive attackers.

Related Works in Same Category

The following **4 sibling papers** share the same taxonomy leaf node with the original paper:

1. Targeted vaccine: Safety alignment for large language models against harmful fine-tuning via layer-wise perturbation

Authors: Liu Guo-zhi, Lin WeiWei, Guozhi Liu, Huang Tiansheng, Weiwei Lin, et al. (12 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Harmful fine-tuning attack poses a serious threat to the online fine-tuning service. Vaccine, a recent alignment-stage defense, applies uniform perturbation to all layers of embedding to make the model robust to the simulated embedding drift. However, applying layer-wise uniform perturbation may lead to excess perturbations for some particular safety-irrelevant layers, resulting in defense performance degradation and unnecessary memory consumption. To address this limitation, we propose Targeted...

Relationship Analysis

Both papers belong to the Perturbation-Based Alignment Enhancement category, applying controlled perturbations during alignment to defend against harmful fine-tuning attacks. They share the core approach of using adversarial perturbations to increase model robustness, with both building upon the Vaccine framework that perturbs embeddings during alignment. The key difference is that the original paper (SEAM) introduces a novel self-destructive mechanism through gradient coupling between benign and harmful tasks with Hessian-free optimization, while the candidate paper (T-Vaccine) focuses on layer-wise selective perturbation using gradient norm-based importance sampling to reduce memory consumption while maintaining defense effectiveness.

---

2. Immunization against harmful fine-tuning attacks

Authors: Rosati, Domenic, Domenic Rosati, Jan Wehner, Kai Williams, et al. (12 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Large Language Models (LLMs) are often trained with safety guards intended to prevent harmful text generation. However, such safety training can be removed by fine-tuning the LLM on harmful datasets. While this emerging threat (harmful fine-tuning attacks) has been characterized by previous work, there is little understanding of how we should proceed in constructing and validating defenses against these attacks especially in the case where defenders would not have control of the fine-tuning proc...

Relationship Analysis

Both papers belong to the Perturbation-Based Alignment Enhancement category, applying controlled perturbations during alignment to increase robustness against harmful fine-tuning. They overlap in their goal of making models resistant to harmful fine-tuning through alignment-stage interventions that modify the training process. However, the original paper (SEAM) focuses on creating self-destructive models that catastrophically degrade under harmful fine-tuning by coupling benign and harmful gradient trajectories, while the candidate paper provides a theoretical framework ('Immunization conditions') for evaluating defenses against harmful fine-tuning attacks without proposing a specific perturbation-based defense mechanism.

---

3. Representation noising: A defence mechanism against harmful finetuning

Authors: David Atanasov, Å□ukasz Bartoszcze, Domenic Rosati, Robie Gonzales, Jan Wehner, et al. (11 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Releasing open-source large language models (LLMs) presents a dual-use risk since bad actors can easily fine-tune these models for harmful purposes. Even without the open release of weights, weight stealing and fine-tuning APIs make closed models vulnerable to harmful fine-tuning attacks (HFAs). While safety measures like preventing jailbreaks and improving safety guardrails are important, such measures can easily be reversed through fine-tuning. In this work, we propose Representation Noising (...

Relationship Analysis

Both papers belong to the Perturbation-Based Alignment Enhancement category, applying controlled perturbations during alignment to defend against harmful fine-tuning attacks. They overlap in their core approach of modifying model representations during alignment to increase robustness, with both using adversarial training components and gradient-based techniques. The key difference is that the original paper (SEAM) focuses on coupling benign and harmful optimization trajectories to create self-destructive models that collapse under attack, while the candidate paper (RepNoise) focuses on removing information about harmful representations by pushing them toward random noise using Maximum Mean Discrepancy and layer-wise gradient ascent.

---

4. Vaccine: Perturbation-aware Alignment for Large Language Models against Harmful Fine-tuning Attack

Authors: Sihao Hu, Tiansheng Huang, Ling Liu | **Year/Venue**: 2024 | **URL**: View paper

Abstract

N/A

Relationship Analysis

Both papers belong to the Perturbation-Based Alignment Enhancement category, applying controlled perturbations during alignment to defend against harmful fine-tuning attacks. They overlap in their core approach of modifying the alignment process to increase robustness against embedding drift caused by adversarial fine-tuning. The key difference is that the original paper (SEAM) introduces a self-destructive mechanism that couples benign and harmful optimization trajectories to cause catastrophic performance collapse under

strong attacks, while Vaccine focuses on perturbation-aware alignment through adversarial training to mitigate embedding shifts without the self-destructive property.

Contributions Analysis

Overall novelty summary. The paper introduces SEAM, a defense mechanism that transforms aligned language models into 'self-destructive' systems that degrade when fine-tuned on harmful data while preserving performance on legitimate tasks. This work resides in the Perturbation-Based Alignment Enhancement leaf, which contains five papers including the original submission. This leaf represents a moderately populated research direction within the broader Alignment-Stage Defense Mechanisms branch, suggesting active but not overcrowded exploration of perturbation-based approaches to harmful fine-tuning defense.

The taxonomy reveals that perturbation-based methods sit alongside three sibling approaches within alignment-stage defenses: gradient-based optimization (four papers), safety data curation (three papers), and tamper-resistant safeguards (two papers). The perturbation-based leaf appears slightly more populated than these alternatives, indicating sustained interest in representation-level interventions. Neighboring branches address orthogonal threat windows—runtime detection mechanisms and post-fine-tuning recovery—while the adversarial training branch (seven papers across three leaves) explores complementary robustness-building strategies that could potentially integrate with alignment-stage defenses.

Among the three contributions analyzed from 30 candidate papers examined, the core SEAM defense method and the novel loss function coupling benign/harmful trajectories show no clear refutation across 10 candidates each. However, the Hessian-free gradient estimation technique encountered three refutable candidates among 10 examined, suggesting this computational component has more substantial prior work in optimization literature. The limited search scope means these statistics reflect top-30 semantic matches rather than exhaustive coverage, and the self-destructive model concept appears less explored than the underlying gradient estimation machinery.

Based on the limited literature search covering 30 candidates, SEAM's core conceptual contribution—intentionally coupling optimization trajectories to induce performance degradation on harmful data—appears relatively novel within the perturbation-based alignment enhancement space. The analysis cannot rule out relevant work outside the top-30 semantic matches or in adjacent optimization subfields. The gradient estimation component shows clearer connections to existing techniques, which is expected for a foundational computational tool adapted to this specific defense context.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: SEAM: Self-destructive language model defense method

Description: The authors propose SEAM, a defense method that transforms large language models into self-destructive models. These models maintain capabilities for legitimate tasks while exhibiting substantial performance degradation when fine-tuned on harmful data, creating intrinsic resistance to harmful fine-tuning attacks.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Safety misalignment against large language models

URL: View paper

Brief Assessment

Safety misalignment against large[61] focuses on attacking safety alignment through fine-tuning methods (supervised and self-supervised), while SEAM proposes a defense mechanism that creates self-destructive models. These are complementary rather than overlapping research directions - one attacks alignment, the other defends it through intrinsic model properties.

---

2. Deep ignorance: Filtering pretraining data builds tamper-resistant safeguards into open-weight llms

URL: View paper

Brief Assessment

Deep ignorance[65] focuses on pretraining data filtering to prevent models from learning harmful knowledge, while SEAM introduces a post-training defense that couples optimization trajectories to cause performance degradation during harmful fine-tuning. These are fundamentally different approaches to model safety.

---

3. Targeted vaccine: Safety alignment for large language models against harmful fine-tuning via layer-wise perturbation

URL: View paper

Brief Assessment

Targeted vaccine[1] focuses on layer-wise perturbation to improve memory efficiency and defense performance against harmful fine-tuning, but does not create self-destructive models that degrade general performance when fine-tuned on harmful data. The mechanisms are fundamentally different: Targeted vaccine[1] selectively applies perturbations to safety-critical layers, while SEAM couples optimization trajectories to cause catastrophic performance collapse.

---

4. Lisa: Lazy Safety Alignment for Large Language Models against Harmful Fine-tuning Attack

URL: View paper

Brief Assessment

Lisa[20] focuses on separating optimization states during fine-tuning to prevent harmful fine-tuning attacks, using a proximal term to stabilize convergence. This differs fundamentally from SEAM's approach of creating self-destructive models that degrade performance when fine-tuned on harmful data through coupled optimization trajectories.

---

5. Defending Against Prompt Injection with DataFilter

URL: View paper

Brief Assessment

Defending Against Prompt Injection[63] addresses prompt injection attacks through data filtering at test-time, while SEAM focuses on harmful fine-tuning attacks through alignment-stage training that creates self-destructive models. These are fundamentally different threat models and defense mechanisms.

---

6. Mitigating fine-tuning based jailbreak attack with backdoor enhanced safety alignment

URL: View paper

Brief Assessment

Mitigating fine-tuning based jailbreak[45] focuses on backdoor-enhanced safety alignment during fine-tuning by adding prefixed safety examples with secret prompts, not on creating self-destructive models that degrade performance when fine-tuned on harmful data.

---

7. Vaccine: Perturbation-aware Alignment for Large Language Models against Harmful Fine-tuning Attack

URL: View paper

Brief Assessment

Vaccine[33] is not available in the provided candidate paper context. Without access to the full text of Vaccine[33], I cannot assess whether it refutes the novelty of SEAM's self-destructive defense mechanism that couples optimization trajectories of benign and harmful data.

---

8. Representation noising: A defence mechanism against harmful finetuning

URL: View paper

Brief Assessment

Representation noising[12] focuses on removing information about harmful representations by pushing them towards random noise, making recovery difficult during fine-tuning. SEAM creates self-destructive models through gradient coupling between benign and harmful tasks, causing catastrophic performance collapse under harmful fine-tuning. These are fundamentally different defense mechanisms with distinct technical approaches.

---

9. Safe Delta: Consistently Preserving Safety when Fine-Tuning LLMs on Diverse Datasets

URL: View paper

Brief Assessment

Safe Delta[62] addresses fine-tuning safety through post-training parameter adjustment (delta parameters), while SEAM creates intrinsic self-destructive properties during alignment training. These are fundamentally different defense mechanisms operating at different stages.

---

10. Fight Fire with Fire: Defending Against Malicious RL Fine-Tuning via Reward Neutralization

URL: View paper

Brief Assessment

Fight Fire with Fire[64] focuses on defending against malicious RL fine-tuning through reward neutralization, while SEAM addresses harmful supervised fine-tuning by creating self-destructive models. These are fundamentally different attack vectors and defense mechanisms.

Contribution 2: Novel loss function coupling benign and harmful optimization trajectories

Description: The authors introduce a novel loss function that deliberately couples the optimization trajectories of harmful and benign tasks. This coupling ensures that attempts to optimize for harmful objectives inevitably lead to degradation in general model performance, enhanced with adversarial gradient ascent.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. From Hallucinations to Jailbreaks: Rethinking the Vulnerability of Large Foundation Models

URL: View paper

Brief Assessment

From Hallucinations to Jailbreaks[73] focuses on the interplay between hallucinations and jailbreaks in vision-language models through attention-level and token-level optimization, not on coupling benign and harmful task trajectories during alignment training to create self-destructive models.

---

2. Lyapunov-based safe policy optimization for continuous control

URL: View paper

Brief Assessment

Lyapunov-based safe policy optimization[66] addresses safe reinforcement learning through Lyapunov constraints for continuous control, not loss functions coupling benign and harmful task trajectories in language model alignment.

---

3. Safe exploration for optimization with Gaussian processes

URL: View paper

Brief Assessment

Safe exploration for optimization[70] addresses safe exploration in Gaussian process optimization for sequential decision problems, not adversarial fine-tuning of language models. The paper focuses on ensuring sampled function values exceed safety thresholds during optimization, which is fundamentally different from coupling harmful and benign task trajectories in LLM alignment.

---

4. Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics

URL: View paper

Brief Assessment

Bayesian optimization with safety[67] addresses safe parameter tuning in robotics with safety constraints on system performance, not adversarial coupling of benign and harmful task gradients in language model fine-tuning.

---

5. Safer Conflict-Based Search: Risk-Constrained Optimal Pathfinding for Multiple Connected and Automated Vehicles

URL: View paper

Brief Assessment

Safer Conflict-Based Search[72] focuses on risk-constrained pathfinding for connected and automated vehicles in physical coordination tasks, not on loss functions for language model alignment or coupling optimization trajectories of benign versus harmful tasks.

---

6. A Closer Look at Smoothness in Domain Adversarial Training

URL: View paper

Brief Assessment

A Closer Look at[68] focuses on domain adversarial training for unsupervised domain adaptation, analyzing smoothness properties of task loss versus adversarial loss components. This is fundamentally different from coupling benign and harmful task optimization trajectories for safety alignment in LLMs.

---

7. Enhanced Route Optimization: Incorporating Road Safety Factors for Optimal Path Selection

URL: View paper

Brief Assessment

Enhanced Route Optimization[75] focuses on route planning algorithms using Dijkstra's algorithm with safety and travel time factors. It does not address loss functions, optimization trajectories, or machine learning model training, making it entirely unrelated to the original paper's contribution on coupling benign and harmful task optimization in language models.

---

8. Safe Value Functions: Learned Critics as Hard Safety Constraints

URL: View paper

Brief Assessment

Safe Value Functions[71] focuses on learning control barrier functions for robotic safety constraints through reinforcement learning value functions, not on coupling optimization trajectories of benign and harmful tasks in language model alignment.

---

9. DGA-ACO: Enhanced Dynamic Genetic Algorithmâ□□Ant Colony Optimization Path Planning for Agribots.

URL: View paper

Brief Assessment

DGA-ACO[69] focuses on agricultural robot path planning using genetic algorithms and ant colony optimization. It addresses trajectory optimization for physical navigation tasks, not loss functions for language model alignment or coupling of benign/harmful optimization trajectories in machine learning contexts.

---

10. A multiobjective optimization algorithm for safety and optimality of 3-D route planning in UAV

URL: View paper

Brief Assessment

A multiobjective optimization algorithm[74] addresses UAV path planning with multiple objectives (fuel cost, safety, smoothness) using the hybrid slime mould algorithm. This is fundamentally different from the original paper's loss function that couples benign and harmful task optimization trajectories in language model alignment.

Contribution 3: Efficient Hessian-free gradient estimate with theoretical error bounds

Description: The authors develop an efficient Hessian-free gradient estimation method that makes the optimization of their loss function computationally tractable for large models. They provide theoretical error bounds (Theorem 1) for this approximation method.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications

URL: View paper

Brief Assessment

A primer on zeroth-order[55] focuses on zeroth-order optimization methods that use only function evaluations without requiring gradients. The original paper develops a Hessian-free gradient estimation method specifically for coupling benign and harmful optimization trajectories in LLM safety alignment, which is a fundamentally different problem domain and technical approach.

---

2. Moreau envelope for nonconvex bi-level optimization: A single-loop and hessian-free solution strategy

URL: View paper

Brief Assessment

Moreau envelope for nonconvex[59] addresses bi-level optimization with Hessian-free methods, not the specific context of harmful fine-tuning defense for LLMs. The technical domains and problem formulations are fundamentally different.

---

3. Hessian aided policy gradient

URL: View paper

Prior Art Analysis

Hessian aided policy gradient[60] demonstrates prior work on Hessian-free gradient estimation methods with theoretical error bounds in the context of policy gradient optimization. The candidate paper presents a finite difference method for computing Hessian-vector products with explicit error bounds, and develops a variance-reduced gradient estimator that utilizes Hessian information without explicitly computing the full Hessian matrix. This work predates the original paper and establishes both the computational tractability and theoretical guarantees for Hessian-free approximations in gradient-based optimization.

Evidence

Evidence 1 - **Rationale**: Both papers provide theoretical error bounds for Hessian-free approximation methods. The candidate explicitly bounds the approximation error of the finite difference method for Hessian-vector products, demonstrating prior work on this theoretical guarantee. - **Original**: we develop an efficient hessian-free gradient estimate with theoretical error bounds - **Candidate**: assuming the second-order smoothness of φ, i.e. ‖∇2φ(x) -∇2φ(y)‖≤ l2‖x -y‖ for arbitrary x,y ∈rd, we bound ‖ξϵ(v; φ) -∇2φ(θ) ·v‖≤ l2‖v‖ϵ, (17) which can be made arbitrarily small by taking a sufficiently small ϵ.

Evidence 2 - **Rationale**: Both papers emphasize computational efficiency of Hessian-free methods. The candidate demonstrates that Hessian-vector products can be computed in linear time O(hd) without explicitly forming the Hessian matrix, establishing prior work on efficient Hessian-free gradient estimation. - **Original**: To enable practical training, we develop an efficient hessian-free gradient estimate with theoretical error bounds. - **Candidate**: let us reemphasize that the estimator (15) can be computed in linear time in terms of the parameter dimension d: first note that computing Δt is equivalent to computing |m| matrix-vector product ˜∇2(θ; τ) ·v. from (14) we can write ˜∇2(θ; τ)·v=(∇log p(τ; πθ)⊤v)∇φ(θ; τ) +∇2φ(θ; τ)·v. clearly, the firs...

---

4. A fully single loop algorithm for bilevel optimization without hessian inverse

URL: View paper

Brief Assessment

A fully single loop[56] focuses on bilevel optimization problems with Hessian-free gradient estimation for hyper-gradient computation, not on the alignment-enhancing defense context of harmful fine-tuning attacks. The technical domains and problem formulations are fundamentally different.

---

5. On the convergence theory of gradient-based model-agnostic meta-learning algorithms

URL: View paper

Prior Art Analysis

On the convergence theory[53] demonstrates prior work on Hessian-free gradient estimation methods with theoretical error bounds in the context of gradient-based model-agnostic meta-learning (MAML). The candidate paper presents a Hessian-free gradient estimate with formal theoretical error bounds (Theorem 1 in their work), predating the original paper's contribution. Both papers develop approximation methods to avoid expensive Hessian computations while providing theoretical guarantees on approximation error.

Evidence

Evidence 1 - **Rationale**: This demonstrates that the candidate paper's Hessian-free method achieves computational efficiency (O(d) instead of O(d²)) while maintaining theoretical guarantees, establishing prior work on efficient Hessian-free gradient estimation with error bounds. - **Original**: we develop an efficient hessian-free gradient estimate with theoretical error bounds. extensive evaluation across llms and datasets demonstrates that seam creates a no-win situation for adversaries - **Candidate**: comparing the results in theorem4.17 for hf-maml with the result in theorem4.12 for maml shows that the complexity of these methods and the resulted accuracy are the same, up to a constant factor. hence, hf-maml recovers the complexity of mamlwithout computing second-order information or performing ...

---

6. Model-agnostic meta-policy optimization via zeroth-order estimation: A linear quadratic regulator perspective

URL: View paper

Prior Art Analysis

Model-agnostic meta-policy optimization via[51] demonstrates prior work on Hessian-free gradient estimation methods with theoretical error bounds in the context of meta-learning for LQR problems. The candidate paper explicitly develops a zeroth-order meta-gradient estimation framework that eliminates Hessian computation and provides theoretical guarantees including error bounds (Lemma 6, Lemma 7). The candidate's approach uses Stein's Gaussian smoothing technique to approximate gradients without computing Hessians, similar to the original paper's goal of making optimization tractable for large models. Both papers address the computational intractability of Hessian estimation and provide theoretical bounds on approximation errors.

Evidence

Evidence 1 - **Rationale**: Both papers explicitly claim to develop Hessian-free gradient estimation methods that address computational tractability for large models. - **Original**: we develop an efficient hessian-free gradient estimate with theoretical error bounds, making seam practical for large models. - **Candidate**: we develop a zeroth-order meta-gradient estimation framework, presented in algorithm 2. this hessian-free approach eliminates the instability and high computational cost associated with exact meta-gradient estimation.

Evidence 2 - **Rationale**: Both papers provide formal theoretical bounds on the approximation error of their Hessian-free gradient estimation methods, demonstrating prior work exists with similar theoretical guarantees. - **Original**: theorem 1. the approximation error of the hessian-free gradient estimate\∇θlsd(θ) is upper bounded by: ‖ \∇θlsd(θ) - ∇θlsd(θ)‖ ≤є lh a ‖ga(θ)‖ + lh b ‖gb(θ)‖ + o є2 - **Candidate**: lemma 6 (gradient estimation). for sufficiently small numbers є, δ∈ (0, 1), given a control policy k, let ℓ, radius r, number of trajectories m satisfying the following dependence... then, with probability at least 1 - 2δ, the gradient estimation error is bounded by ‖∇ji(k) - ˜∇ji(k)‖f ≤ є

Evidence 3 - **Rationale**: The candidate provides theoretical error bounds (Lemma 7) for meta-gradient estimation without Hessian computation, demonstrating that similar work with theoretical guarantees existed prior to the original paper's submission. - **Original**: to enable practical training, we develop an efficient hessian-free gradient estimate with theoretical error bounds. - **Candidate**: lemma 7 (meta-gradient estimation). for sufficiently small numbers є, δ∈ (0, 1), given a control policy k... then, for each iteration the meta-gradient estimation is є-accurate, i.e., ‖˜∇l(k) - ∇l(k)‖f ≤ є with probability at least 1 - δ.

Evidence 4 - **Rationale**: Both papers identify the computational intractability of Hessian-based methods and propose Hessian-free alternatives as solutions to this problem. - **Original**: while directly optimizing this formulation is computationally intractable, we develop an efficient hessian-free gradient estimate with theoretical error bounds - **Candidate**: the complete maml policy gradient methods for such a meta-objective require differentiating through the optimization process, which necessitates the estimation of hessians or even higher order information, making them computationally expensive and unstable... this incentivizes us to focus our attent...

---

7. Optimal Hessian/Jacobian-free nonconvex-PL bilevel optimization

URL: View paper

Brief Assessment

Optimal HessianJacobian-free nonconvex-PL bilevel[54] addresses bilevel optimization problems with Hessian-free methods, not the alignment-enhancing defense mechanisms for LLMs that the original paper focuses on. The technical contexts are fundamentally different.

---

8. Achieving Complexity in Hessian/Jacobian-free Stochastic Bilevel Optimization

URL: View paper

Brief Assessment

Achieving Complexity in HessianJacobian-free[58] focuses on bilevel optimization with Hessian/Jacobian-free methods, not on language model alignment or harmful fine-tuning defense. The technical contexts are fundamentally different.

---

9. New insights and perspectives on the natural gradient method

URL: View paper

Brief Assessment

New insights and perspectives[52] focuses on natural gradient methods and 2nd-order optimization for general machine learning models, not specifically on harmful fine-tuning defense or LLM safety alignment. The Hessian-free techniques discussed are applied to different optimization contexts (natural gradient descent, Fisher information matrices) rather than the specific loss function coupling mechanism described in the original paper.

---

10. Robust dataâ□□driven dynamic optimization using a setâ□□based gradient estimator

URL: View paper

Brief Assessment

The candidate paper focuses on extremum-seeking control with a polyhedral set-based gradient estimator for nonlinear systems, not on Hessian-free gradient estimation methods for language model optimization. These are fundamentally different technical domains and problem settings.

Appendix: Text Similarity Detection

Textual similarity detection checked 31 papers and found 1 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Targeted vaccine: Safety alignment for large language models against harmful fine-tuning via layer-wise perturbation

Detected in: Core Task (sibling), Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Self-Destructive Language Models View paper
- [1] Targeted vaccine: Safety alignment for large language models against harmful fine-tuning via layer-wise perturbation View paper
- [2] Safeguarding large language models in real-time with tunable safety-performance trade-offs View paper
- [3] Open-weight genome language model safeguards: Assessing robustness via adversarial fine-tuning View paper
- [4] Fine-tuning aligned language models compromises safety, even when users do not intend to! View paper
- [5] Immunization against harmful fine-tuning attacks View paper
- [6] Jailbreak attacks and defenses against large language models: A survey View paper
- [7] Securing large language models: Threats, vulnerabilities and responsible practices View paper
- [8] A survey on backdoor threats in large language models (llms): Attacks, defenses, and evaluations View paper
- [9] Defending against alignment-breaking attacks via robustly aligned llm View paper
- [10] Sdd: Self-degraded defense against malicious fine-tuning View paper
- [11] Scam Shield: Multi-Model Voting and Fine-Tuned LLMs Against Adversarial Attacks View paper
- [12] Representation noising: A defence mechanism against harmful finetuning View paper
- [13] Adversarial attacks and defenses for large language models (LLMs): methods, frameworks & challenges View paper
- [14] Towards understanding the fragility of multilingual llms against fine-tuning attacks View paper
- [15] Helping Large Language Models Protect Themselves: An Enhanced Filtering and Summarization System View paper
- [16] Adaptive defense against harmful fine-tuning for large language models via bayesian data scheduler View paper
- [17] Securing vision-language models with a robust encoder against jailbreak and adversarial attacks View paper
- [18] Safe lora: The silver lining of reducing safety risks when finetuning large language models View paper
- [19] On weaponization-resistant large language models with prospect theoretic alignment View paper
- [20] Lisa: Lazy Safety Alignment for Large Language Models against Harmful Fine-tuning Attack View paper
- [21] Safety fine-tuning at (almost) no cost: A baseline for vision large language models View paper
- [22] Threat Landscape of Adversarial Attacks on Generative AI and Large Language Models (LLMs): Exploring Different Types of Adversarial Attacks, Associated Risks â¦ View paper
- [23] Baseline defenses for adversarial attacks against aligned language models View paper
- [24] Gradient surgery for safe llm fine-tuning View paper
- [25] Unified Defense for Large Language Models against Jailbreak and Fine-Tuning Attacks in Education View paper
- [26] Robust safety classifier for large language models: Adversarial prompt shield View paper
- [27] Dual-Use of Large Language Models (LLMs) and Generative AI (GenAI) in Cybersecurity: Risks, Defenses, and Governance Strategies View paper
- [28] Probe before you talk: Towards black-box defense against backdoor unalignment for large language models View paper
- [29] Adversarial training for multimodal large language models against jailbreak attacks View paper
- [30] MetaDefense: Defending Finetuning-based Jailbreak Attack Before and During Generation View paper
- [31] How should pre-trained language models be fine-tuned towards adversarial robustness? View paper
- [32] Adversarial Attacks on Large Language Modelâ□□Based System and Mitigating Strategies: A Case Study on ChatGPT View paper
- [33] Vaccine: Perturbation-aware Alignment for Large Language Models against Harmful Fine-tuning Attack View paper
- [34] Large language model sentinel: Llm agent for adversarial purification View paper
- [35] Lazy Safety Alignment for Large Language Models against Harmful Fine-tuning View paper
- [36] Safety at scale: A comprehensive survey of large model safety View paper
- [37] Adversarial tuning: Defending against jailbreak attacks for llms View paper
- [38] Harmful fine-tuning attacks and defenses for large language models: A survey View paper
- [39] Analysing Safety Risks in LLMs Fine-Tuned with Pseudo-Malicious Cyber Security Data View paper
- [40] Backdooralign: Mitigating fine-tuning based jailbreak attack with backdoor enhanced safety alignment View paper
- [41] Pharmacist: Safety Alignment Data Curation for Large Language Models against Harmful Fine-tuning View paper
- [42] Keeping llms aligned after fine-tuning: The crucial role of prompt templates View paper
- [43] SmoothLLM: Defending Large Language Models Against Jailbreaking Attacks View paper
- [44] Attack prompt generation for red teaming and defending large language models View paper
- [45] Mitigating fine-tuning based jailbreak attack with backdoor enhanced safety alignment View paper
- [46] Mitigating fine-tuning risks in llms via safety-aware probing optimization View paper
- [47] Self-evaluation as a defense against adversarial attacks on llms View paper
- [48] Large Language Models (LLMs) and Generative AI in Cybersecurity and Privacy: A Survey of Dual-Use Risks, AI-Generated Malware, Explainability, and Defensive â¦ View paper
- [49] Tamper-resistant safeguards for open-weight llms View paper
- [50] Ethical Treatment of Language Models Against Harmful Inference-Time Interventions View paper
- [51] Model-agnostic meta-policy optimization via zeroth-order estimation: A linear quadratic regulator perspective View paper
- [52] New insights and perspectives on the natural gradient method View paper
- [53] On the convergence theory of gradient-based model-agnostic meta-learning algorithms View paper
- [54] Optimal Hessian/Jacobian-free nonconvex-PL bilevel optimization View paper
- [55] A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications View paper
- [56] A fully single loop algorithm for bilevel optimization without hessian inverse View paper
- [57] Robust dataâ□□driven dynamic optimization using a setâ□□based gradient estimator View paper

- [58] Achieving Complexity in Hessian/Jacobian-free Stochastic Bilevel Optimization View paper
- [59] Moreau envelope for nonconvex bi-level optimization: A single-loop and hessian-free solution strategy View paper
- [60] Hessian aided policy gradient View paper
- [61] Safety misalignment against large language models View paper
- [62] Safe Delta: Consistently Preserving Safety when Fine-Tuning LLMs on Diverse Datasets View paper
- [63] Defending Against Prompt Injection with DataFilter View paper
- [64] Fight Fire with Fire: Defending Against Malicious RL Fine-Tuning via Reward Neutralization View paper
- [65] Deep ignorance: Filtering pretraining data builds tamper-resistant safeguards into open-weight llms View paper
- [66] Lyapunov-based safe policy optimization for continuous control View paper
- [67] Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics View paper
- [68] A Closer Look at Smoothness in Domain Adversarial Training View paper
- [69] DGA-ACO: Enhanced Dynamic Genetic Algorithmâ□□Ant Colony Optimization Path Planning for Agribots. View paper
- [70] Safe exploration for optimization with Gaussian processes View paper
- [71] Safe Value Functions: Learned Critics as Hard Safety Constraints View paper
- [72] Safer Conflict-Based Search: Risk-Constrained Optimal Pathfinding for Multiple Connected and Automated Vehicles View paper
- [73] From Hallucinations to Jailbreaks: Rethinking the Vulnerability of Large Foundation Models View paper
- [74] A multiobjective optimization algorithm for safety and optimality of 3-D route planning in UAV View paper
- [75] Enhanced Route Optimization: Incorporating Road Safety Factors for Optimal Path Selection View paper