Machine Learning for Language Processing
Dr. Dietrich Klakow

Seminar Report
January 2026

Overtrained Language Models Are Harder to Fine-Tune

Based on the work by

Jacob Mitchell Springer, Sachin Goyal, Kaiyue Wen,
Tanishq Kumar, Xiang Yue, Sadhika Malladi,
Graham Neubig, Aditi Raghunathan

Code Repository
https://github.com/Abdullahw72

Abdullah (7075730)
abyy00002@stud.uni-saarland.de

Abstract

The dominant paradigm in large language model (LLM) development assumes that increasing pre-training data monotonically improves downstream performance. However, recent work by Springer et al. (2025) challenges this assumption by identifying *catastrophic overtraining*: a regime in which extended pre-training degrades a model's adaptability to fine-tuning and parameter perturbations.

This report presents a detailed reproduction and analysis of the paper's core empirical and theoretical findings using the Pythia-70M model family. Through four experiments; Gaussian parameter perturbations, fixed learning rate fine-tuning, learning rate trade-off analysis, and a two-layer linear model simulation, we demonstrate that extended pre-training induces progressive sensitivity to parameter updates. While base model perplexity continues to improve, overtrained checkpoints exhibit severe degradation in out-of-distribution (OOD) performance after both random and task-driven modifications.

Our results qualitatively and quantitatively confirm the existence of an inflection point beyond which additional pre-training becomes detrimental. These findings highlight a fundamental trade-off between pre-training optimality and post-training robustness, calling into question the prevailing "more is better" scaling philosophy.

1 Introduction

Large Language Models (LLMs) are typically pre-trained on massive corpora comprising hundreds of billions or even trillions of tokens. This practice is motivated by empirical scaling laws that associate increased data and compute with improved performance [Kaplan et al., 2020]. The success of models such as GPT-4 and LLaMA-3 has reinforced the belief that extended pre-training universally benefits downstream performance.

However, recent evidence suggests that this assumption may be incomplete. While additional pre-training improves base model perplexity and zero-shot performance, it may simultaneously reduce a model's ability to adapt to new tasks. Models are now routinely trained far beyond compute-optimal regimes identified by recent scaling analyses [Hoffmann et al., 2022]. Springer et al. (2025) formalize this tension through the concept of *catastrophic overtraining*.

The paper introduces two central concepts. First, *progressive sensitivity* describes the monotonic increase in performance degradation caused by fixed-magnitude parameter perturbations as pre-training progresses. Second, *catastrophic overtraining* occurs when this sensitivity growth overtakes the base model improvements from additional pre-training, leading to worse post-training performance despite better base perplexity.

This report reproduces and analyzes these claims using the Pythia-70M model suite. Unlike full-scale replications conducted in the original paper, our goal is a mechanistic and qualitative validation under realistic compute constraints. By grounding each experiment in both empirical measurements and theoretical analysis, we aim to clarify why overtraining fundamentally alters the geometry of the optimization landscape.

We reproduce these phenomena across checkpoints spanning 2B–300B tokens and across both random and task-driven parameter updates

2 Related Work

Scaling laws for neural language models suggest predictable improvements in loss as a function of model size, dataset size, and compute [Kaplan et al., 2020]. The Chinchilla scaling law refined this perspective by identifying compute-optimal token-to-parameter ratios, arguing that many contemporary models are significantly undertrained relative to their size [Hoffmann et al., 2022].

Despite these prescriptions, modern foundation models are routinely trained far beyond Chinchilla-optimal regimes. While such overtraining improves pre-training perplexity, its impact on fine-tuning robustness has received comparatively less attention. Classical work on catastrophic forgetting focuses on interference between sequential tasks [McCloskey and Cohen, 1989], but does not explain why sensitivity emerges even under random perturbations.

Recent work on loss landscape sharpness provides a complementary perspective. [Keskar et al., 2017] link sharp minima to reduced generalization and robustness. Springer et al. (2025) extend this intuition by showing that overtrained LLMs occupy increasingly narrow basins of stability, making them fragile to both structured and unstructured updates.

Finally, large-scale training suites such as Pythia provide uniquely controlled checkpoints across training time, enabling mechanistic analysis of training dynamics [Biderman et al., 2023].

3 Methodology

3.1 Model Architecture and Training Regime

All experiments are conducted using the `EleutherAI/pythia-70m` model from the Pythia Scaling Suite [Biderman et al., 2023]. Pythia-70M is a decoder-only transformer with 6 layers, 8 attention heads per layer, and a hidden dimension of 512, totaling approximately 18.9 million non-embedding parameters. The model is trained on the Pile dataset using a fixed batch size of 2,097,152 tokens per optimization step.

A defining feature of the Pythia suite is the availability of uniformly spaced intermediate checkpoints saved throughout pre-training. This enables controlled analysis of training-time effects without confounding architectural or optimization changes. We evaluate checkpoints ranging from `step1000` (approximately 2B tokens) to `step143000` (approximately 300B tokens), corresponding to token-to-parameter ratios spanning from $\sim 10^2$ to $\sim 1.6 \times 10^4$, placing later checkpoints deep in the overtrained regime relative to Chinchilla-optimal scaling.

3.2 Evaluation Tasks

We evaluate model robustness along two axes:

Out-of-Domain (OOD) Retention OOD performance is measured using perplexity on the WikiText-2 test set. WikiText-2 is sufficiently distinct from the Pile to act as a probe of general language modeling ability rather than memorization of the pre-training corpus.

In-Domain (ID) Adaptation ID performance is measured using accuracy on the SST-2 sentiment classification task. This task requires only modest linguistic reasoning but demands structured parameter updates during fine-tuning, making it well-suited for probing adaptation–retention trade-offs.

3.3 Perturbation Protocol

To isolate intrinsic parameter sensitivity, we apply unstructured Gaussian perturbations following Appendix D.2 of Springer et al. (2025). For each parameter tensor θ_i, noise is injected according to:

$$\theta_i' = \theta_i + \gamma \cdot \sigma(\theta_i) \cdot \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I)$$

Scaling noise by the empirical parameter standard deviation ensures that perturbations represent a fixed *relative* displacement across checkpoints. This normalization is critical: using absolute noise magnitudes would artificially inflate sensitivity in later checkpoints purely due to parameter scale differences.

4 Experiment I: Progressive Sensitivity to Gaussian Perturbations

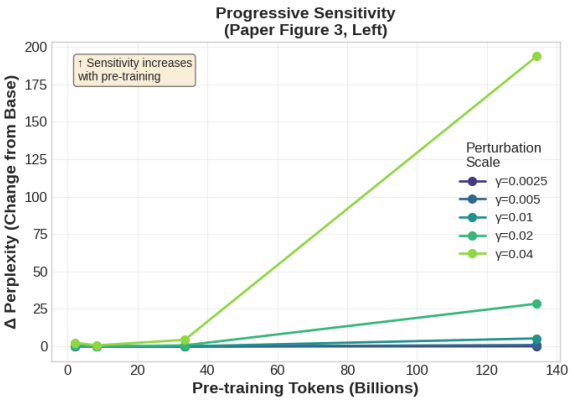
4.1 Objective

The first experiment reproduces Figure 3 of Springer et al. (2025). Its objective is to isolate the intrinsic sensitivity of model parameters by measuring the effect of unstructured Gaussian perturbations on pre-

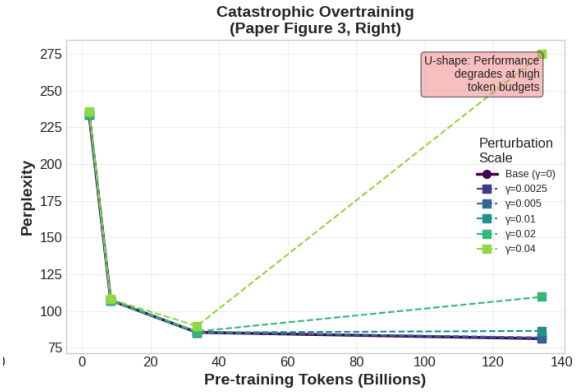
training performance, independent of any downstream task or gradient alignment.

4.2 Results

Figure 1 summarizes the results of applying fixed-magnitude Gaussian perturbations across increasing pre-training budgets.



(a) Progressive sensitivity: change in WikiText-2 perplexity (ΔPPL) relative to the unperturbed base model.



(b) Catastrophic overtraining: absolute WikiText-2 perplexity after perturbation, showing a U-shaped degradation curve.

Figure 1: Effect of Gaussian parameter perturbations across pre-training checkpoints. Left: progressive sensitivity. Right: catastrophic overtraining.

Across all perturbation scales, the change in perplexity increases monotonically with pre-training. For example, at $\gamma = 0.04$, Δ perplexity grows from +2.46 at approximately 2B tokens to +193.86 at 134B tokens, representing nearly a 79-fold increase in sensitivity. This provides direct empirical confirmation of *progressive sensitivity*.

Furthermore, while the base model perplexity continues to improve throughout training, the absolute perplexity of perturbed models begins to increase beyond approximately 33.6B tokens. This divergence marks the onset of *catastrophic overtraining*, where additional pre-training produces models that are strictly worse under even minor parameter perturbations.

4.3 Analysis

Early in training, improvements in base model quality dominate, and the model occupies relatively broad, flat regions of the loss landscape. As training progresses, the model internalizes increasingly fine-grained and fragile features, moving into sharper regions characterized by narrow basins of stability.

Eventually, the rate of sensitivity growth overtakes the rate of base performance improvement. At this point, any fixed-magnitude parameter update, whether random or task-driven, induces net performance degradation, despite continued improvements in unperturbed perplexity.

5 Experiment II: Fine-Tuning with Fixed Learning Rate

5.1 Objective

This experiment evaluates whether progressive sensitivity manifests under structured parameter updates induced by fine-tuning, reproducing Figure 2 of Springer et al. (2025). Unlike Experiment I, which applies random perturbations, this setting reflects realistic downstream adaptation.

5.2 Experimental Setup

All checkpoints are fine-tuned on the SST-2 sentiment classification task for three epochs using a fixed learning rate of 2×10^{-5}. Performance is evaluated along two axes: in-domain (ID) task accuracy and out-of-domain (OOD) knowledge retention measured by WikiText-2 perplexity after fine-tuning.

5.3 Results

Table 1 reports SST-2 accuracy and OOD perplexity after fine-tuning across pre-training checkpoints.

Pre-training Tokens	SST-2 Accuracy	OOD PPL After FT	Δ PPL
2B	0.736	288.47	+42.71
8B	0.787	127.78	+16.71
33B	0.790	102.37	+13.55
134B	0.805	115.19	+29.80
300B	0.802	391.95	+305.39

Table 1: Fine-tuning performance on SST-2 using a fixed learning rate (2×10^{-5}).

5.4 Divergence Between Adaptation and Retention

Figures 2 and 3 reveal a pronounced decoupling between task adaptation and knowledge retention. While SST-2 accuracy improves steadily with pre-training and reaches approximately 80% for later checkpoints, OOD performance deteriorates sharply in the overtrained regime.

Notably, OOD forgetting follows a non-monotonic trajectory, reaching a minimum near 33B tokens before exploding to over +300 perplexity at 300B tokens. This mirrors the inflection point observed in Experiment I, suggesting a shared underlying mechanism.

5.5 Why Fixed Learning Rates Fail

A fixed learning rate implicitly assumes a stable loss landscape across checkpoints. However, as Experiment I demonstrates, extended pre-training drives models into progressively sharper regions of parameter space. Consequently, gradient updates that are benign for early checkpoints become increasingly destructive for overtrained models.

Crucially, the overtrained model does not fail to learn the downstream task. Instead, it succeeds at the cost of overwriting pre-training features, indicating that catastrophic overtraining reflects a failure of robustness rather than a limitation of capacity.

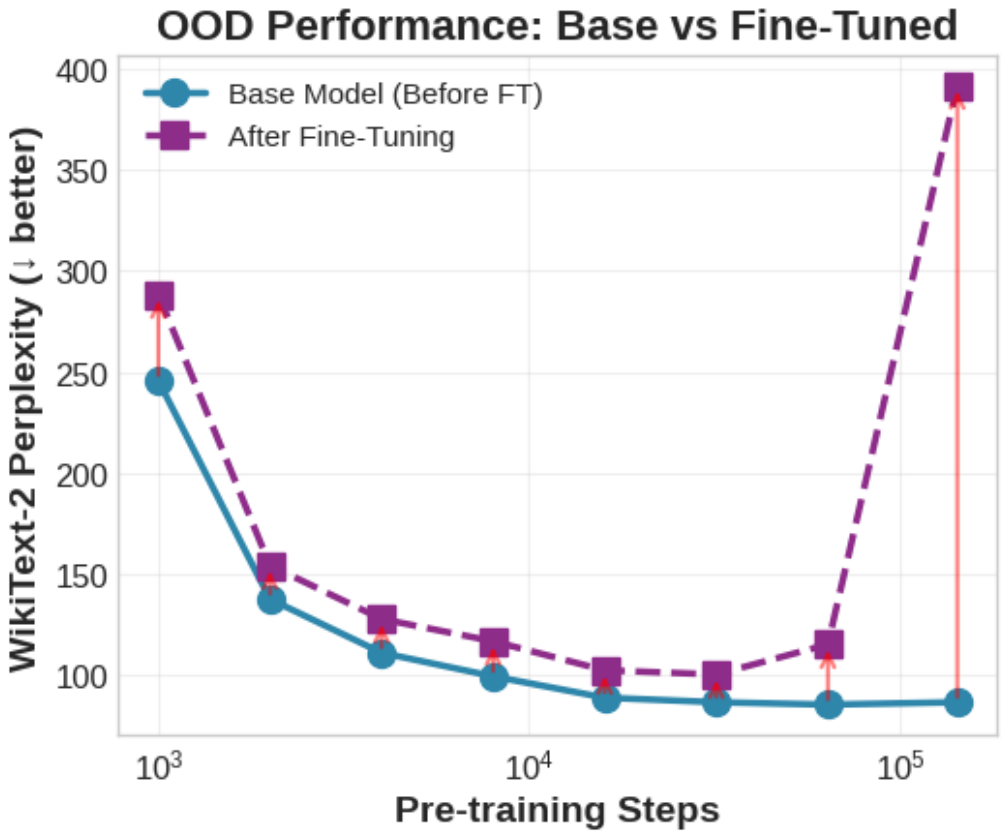
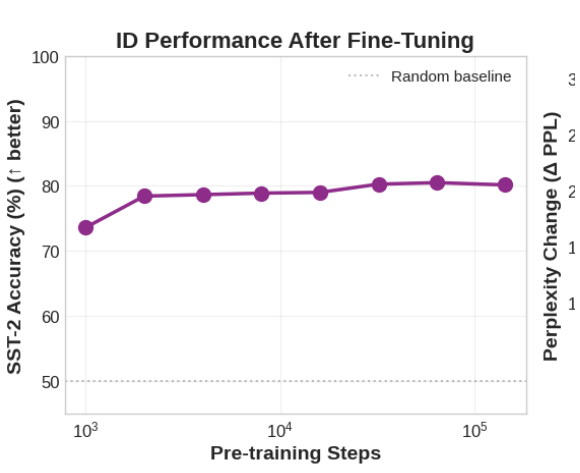
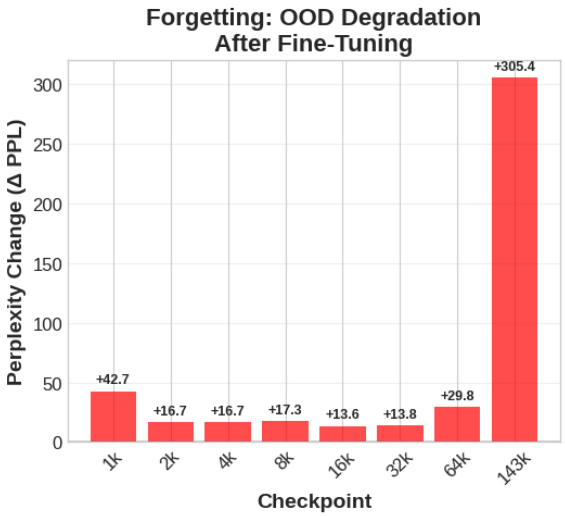


Figure 2: Out-of-domain (OOD) performance on WikiText-2 before and after fine-tuning with a fixed learning rate. While base model perplexity improves monotonically with pre-training, fine-tuned models exhibit severe degradation in the overtrained regime.



(a) OOD forgetting measured as the change in WikiText-2 perplexity after fine-tuning. Forgetting increases sharply for overtrained checkpoints, peaking at the final model.



(b) In-domain (ID) SST-2 accuracy after fine-tuning. Performance improves or saturates with pre-training despite increasing OOD degradation.

Figure 3: Divergence between adaptation and retention under fixed learning-rate fine-tuning. Overtrained models maintain strong ID performance while suffering catastrophic OOD forgetting.

6 Experiment III: Learning Rate Trade-Off

6.1 Objective

This experiment investigates whether learning rate tuning can mitigate catastrophic overtraining.

6.2 Results

Learning rate sweeps reveal a clear trade-off: higher learning rates improve ID accuracy but cause exponentially greater OOD degradation for overtrained checkpoints. Even learning rates that maximize ID performance result in catastrophic forgetting at later stages.

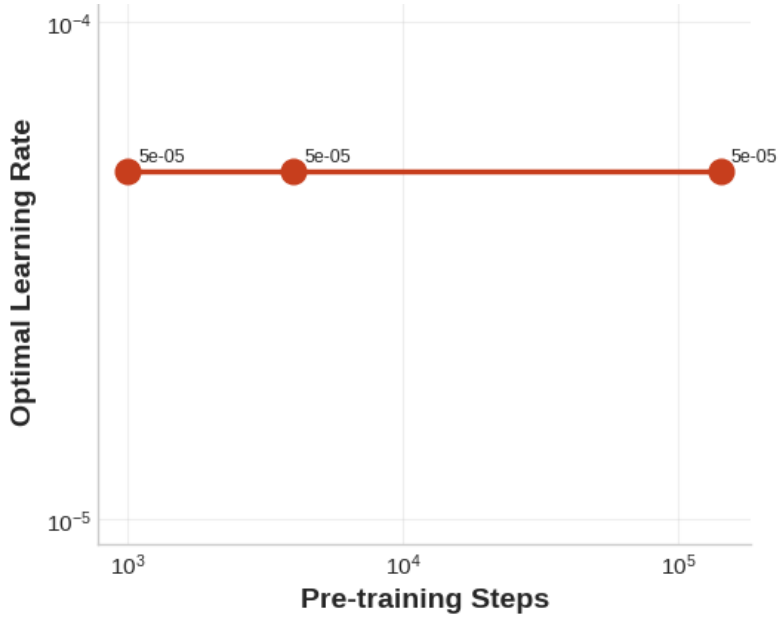


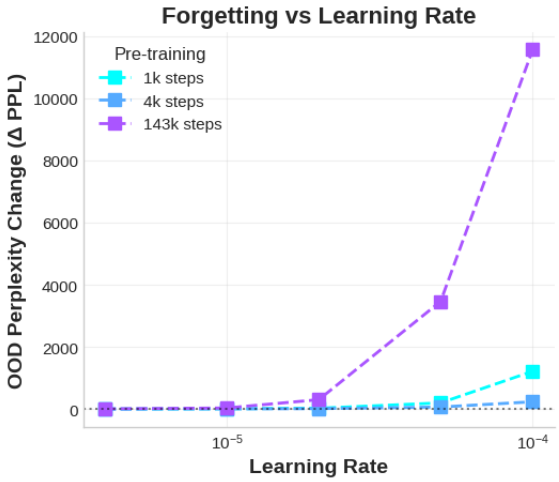
Figure 4: Optimal learning rate for SST-2 fine-tuning as a function of pre-training progress. The optimal learning rate remains approximately constant across checkpoints, indicating that later performance degradation cannot be mitigated by simple learning rate reduction.

Figure 4 shows that the learning rate which maximizes SST-2 accuracy remains approximately constant across pre-training checkpoints. This places Pythia-70M in the "constant optimal learning rate" regime described by Springer et al., where catastrophic overtraining is most pronounced.

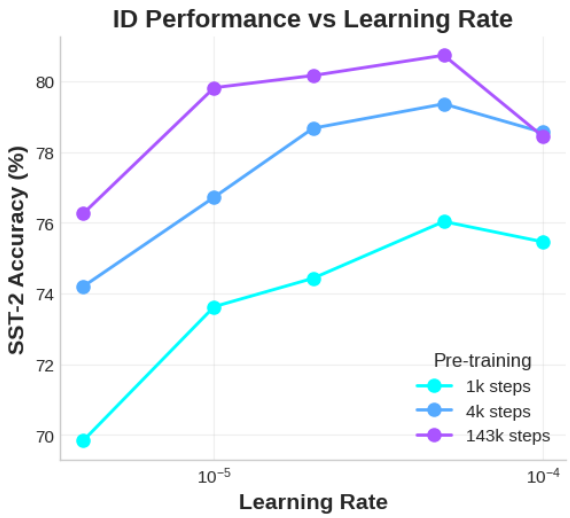
Figures 5a and 5b illustrate the resulting trade-off. While higher learning rates improve in-domain performance, they induce exponentially larger out-of-domain degradation for overtrained models, confirming that learning rate tuning can shift but not eliminate catastrophic overtraining.

6.3 Implications

These findings indicate that learning rate tuning can shift but not eliminate catastrophic overtraining. Even learning rates that are optimal for task accuracy induce catastrophic forgetting in overtrained models, confirming that the underlying issue is structural rather than hyperparameter-specific.



(a) Out-of-domain (OOD) forgetting as a function of learning rate for different pre-training stages. Overtrained models exhibit exponentially greater degradation for the same learning rate, indicating heightened sensitivity.



(b) In-domain (ID) SST-2 accuracy as a function of learning rate for different pre-training stages. Higher learning rates improve task performance but do not mitigate OOD degradation in overtrained models.

Figure 5: Learning rate trade-off under progressive overtraining. While increasing the learning rate improves in-domain performance, it causes disproportionately large out-of-domain forgetting for overtrained checkpoints, revealing a worsening adaptation–retention Pareto frontier.

6.4 Discussion

These findings confirm that learning rate tuning can shift but not eliminate the fundamental trade-off between adaptation and retention. Sensitivity is intrinsic to the parameter configuration induced by overtraining.

7 Experiment IV: Linear Model Analysis of Progressive Sensitivity

7.1 Motivation and Theoretical Background

While Experiments I–III establish catastrophic overtraining empirically in large language models, they do not by themselves explain *why* extended pre-training induces such fragility. To provide mechanistic insight, Springer et al. analyze a two-layer linear network trained via gradient flow, a setting that admits precise characterization of feature learning and parameter sensitivity [Saxe et al., 2019].

In this framework, pre-training corresponds to learning features sequentially according to their singular values, a phenomenon well-characterized in linear networks trained with gradient-based dynamics [Saxe et al., 2019, Gidel et al., 2019]. Early training captures high-variance, robust directions, while later training internalizes increasingly small and fragile features. This mirrors empirical observations in deep networks, where coarse linguistic patterns are learned early and fine-grained correlations emerge late in training.

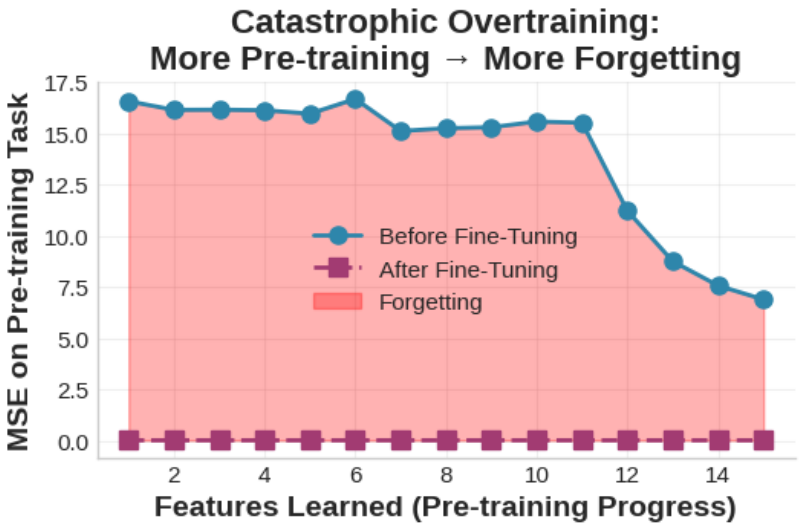


Figure 6: Catastrophic overtraining in a two-layer linear model. As progressively more features are learned during pre-training, fine-tuning on a misaligned task causes increasing degradation in pre-training MSE. Later stages exhibit substantially larger forgetting despite similar fine-tuning success.

7.2 Simulation Setup

We implement the two-layer linear model described in Section 4 of Springer et al. (2025). The model is trained on a synthetic pre-training task until progressively more features are learned. Sensitivity is evaluated under two types of modifications: (i) misaligned fine-tuning on a downstream task, and (ii) unstructured Gaussian parameter perturbations. Performance is measured using mean squared error (MSE) on the original pre-training task.

7.3 Fine-Tuning-Induced Forgetting

Figure 6 shows pre-training task performance before and after fine-tuning as a function of the number of learned features. While early-stage models exhibit limited forgetting, later-stage models suffer substantial performance degradation. This behavior directly parallels Experiment II, where overtrained LLM checkpoints retain task-learning ability but catastrophically lose pre-training knowledge.

7.4 Progressive Sensitivity to Parameter Perturbations

Figure 7 demonstrates that sensitivity to Gaussian perturbations increases monotonically with pre-training progress. Crucially, this effect is observed even though the perturbations are task-agnostic, confirming that fragility arises from the parameter configuration itself rather than task interference.

This result provides a mechanistic explanation for Experiment I: as models internalize smaller singular-value features, fixed-magnitude perturbations disproportionately disrupt model behavior [Gidel et al., 2019].

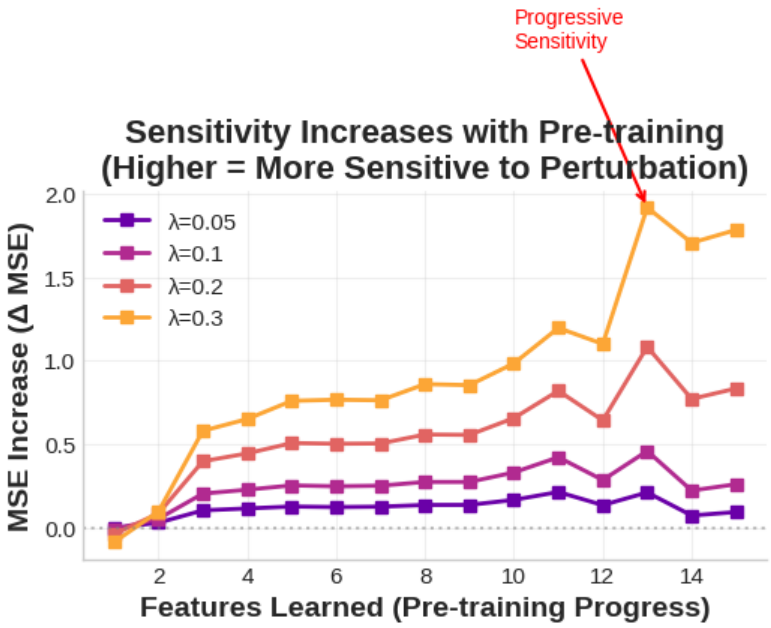


Figure 7: Progressive sensitivity in the linear model. The increase in pre-training MSE (ΔMSE) after Gaussian perturbation grows monotonically as more features are learned. Higher perturbation magnitudes exacerbate this effect, illustrating increasing fragility of later-learned features.

7.5 Absolute Performance After Perturbation

Figure 8 illustrates the emergence of catastrophic overtraining in absolute terms. Although the unperturbed model continues to improve with additional feature learning, the perturbed model's performance degrades beyond a critical point. This reproduces the U-shaped behavior observed in Experiment I and formally derived in Theorem 4.4 of Springer et al. (2025).

7.6 Interpretation and Connection to Deep Models

Together, these results show that catastrophic overtraining is not an artifact of model scale, architecture, or dataset complexity. Instead, it arises from a fundamental property of gradient-based learning: incremental acquisition of increasingly fragile features. As training progresses, models converge to sharper regions of the loss landscape, making them inherently sensitive to both random perturbations and structured fine-tuning updates.

This linear model analysis provides a unifying explanation for all prior experiments, demonstrating that progressive sensitivity is a mechanistic consequence of overtraining rather than a contingent empirical phenomenon.

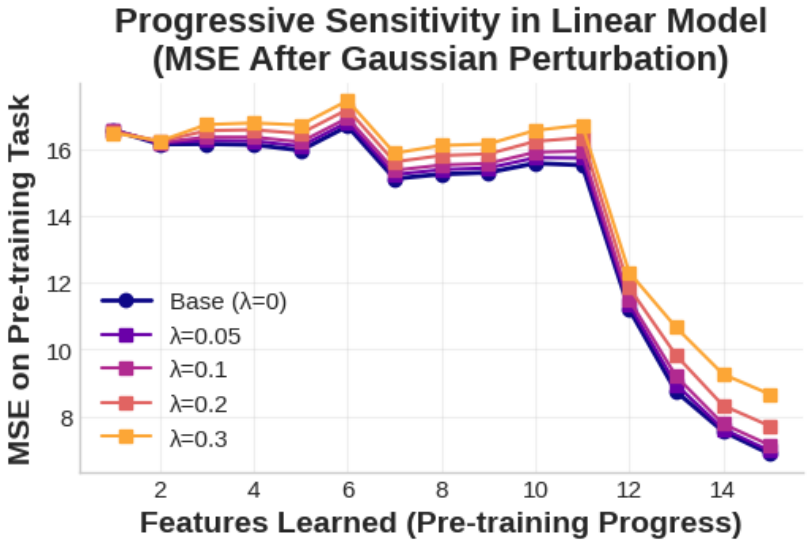


Figure 8: Absolute pre-training MSE after Gaussian perturbation in the linear model. While base performance improves as more features are learned, post-perturbation performance eventually degrades, producing a U-shaped curve characteristic of catastrophic overtraining.

8 Discussion and Conclusion

Across all experiments, a consistent pattern emerges: extended pre-training sharpens the loss landscape, making models increasingly brittle to parameter updates. This brittleness is observable under random noise, fine-tuning, and theoretical analysis, indicating a fundamental property of overtrained models.

This report reproduced and validated the central claims of Springer et al. (2025) using the Pythia-70M model family under realistic compute constraints. Across four complementary experiments, we demonstrated that extended pre-training induces progressive sensitivity to parameter updates, ultimately undermining robustness to both random perturbations and downstream fine-tuning.

Empirically, we showed that while base model perplexity continues to improve with additional pre-training, overtrained checkpoints suffer catastrophic degradation in out-of-domain performance after even modest parameter updates. This effect persists under structured fine-tuning, learning rate tuning, and unstructured noise, indicating that catastrophic overtraining reflects an intrinsic property of the learned parameter configuration rather than a failure of optimization or task mismatch.

The linear model analysis provides a mechanistic explanation for these observations. As models incrementally acquire lower singular-value features, they converge to increasingly sharp regions of the loss landscape. In this regime, fixed-magnitude updates disproportionately disrupt later-learned features, producing the observed trade-off between adaptation and retention.

Taken together, these findings challenge the prevailing assumption that additional pre-training is universally beneficial. Instead, they highlight a fundamental tension between pre-training optimality and post-training robustness. Future work should explore training strategies that preserve plasticity; such as early stopping, regularization, or adaptive optimization, while retaining the benefits of large-scale pre-training [Springer et al., 2025].

9 Limitations and Future Work

Our study is limited to a single model family and a restricted set of downstream tasks. While the qualitative trends closely match those reported by Springer et al. (2025), larger models and more diverse tasks may exhibit additional nuances. Moreover, we did not explore interventions designed to mitigate catastrophic overtraining.

Future work could investigate whether techniques such as adaptive learning rates, regularization during late-stage pre-training, or architectural modifications can preserve robustness without sacrificing base performance. Understanding how to balance scaling with plasticity remains an open and important challenge for large language model development.

References

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning (ICML)*, 2023.

Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Jordan Hoffmann et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Jared Kaplan et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Nitish Keskar et al. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2017.

Michael McCloskey and Neal Cohen. Catastrophic interference in connectionist networks. *Psychology of Learning and Motivation*, 1989.

Andrew M Saxe, James L McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23): 11537–11546, 2019.

Jacob Mitchell Springer, Sachin Goyal, Kaiyue Wen, Tanishq Kumar, Xiang Yue, Sadhika Malladi, Graham Neubig, and Aditi Raghunathan. Overtrained language models are harder to fine-tune. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.