# Robustness to Noisy Labels in Parameter Efficient Fine-tuning

Anonymous ACL submission

## Abstract

As language models grow in size, Parameter Efficient Fine-tuning (PEFT) methods like Low-Rank Adaptation (LoRA) offer compute efficiency while maintaining performance. However, their robustness to label noise, a significant issue in real-world data, remains unexplored. This study investigates whether LoRA-tuned models demonstrate the same level of noise resistance observed in fully fine-tuned Transformer models. Our investigation has multiple key findings: First, we show that LoRA exhibits robustness to random noise similar to full fine-tuning on balanced data, but unlike full fine-tuning, LoRA does not overfit the noisy data. Second, we observe that compared to full fine-tuning, LoRA forgets significantly fewer data points as noise increases. Third, studying how these robustness patterns change as training data becomes imbalanced, we observe that Transformers struggle with imbalanced data, with robustness declining as imbalance worsens. This study highlights LoRA's promise in real-world settings with noise and data imbalance. Overall, our findings reveal LoRA as a robust and efficient alternative for fine-tuning, shedding light on its distinctive characteristics.

## 1 Introduction

In recent years, natural language processing has been revolutionized by large pre-trained language models such as Llama (Touvron et al., 2023), GPT-4 (Achiam et al., 2023), and Gemini (GeminiTeam et al., 2023). However, the massive parameter size of these models, often in the hundreds of millions or billions, presents challenges for fine-tuning and deployment. Parameter Efficient fine-tuning (PEFT) Methods like Low-Rank Adaptation (LoRA; Hu et al., 2022) have emerged as an efficient approach to adapt only a small subset of a large model's parameters for a downstream task (Fu et al., 2023; He et al., 2021). While computationally appealing, it remains unclear whether these parameter-efficient methods exhibit the same characteristics and capabilities as full fine-tuning, especially in terms of robustness to label noise.

Machine learning datasets often contain *label noise*, which occurs when assigned labels to a data point differ from the ground truth. In fact, real-world datasets have been estimated to contain anywhere from 8.0% to 38.5% of noisy labels (Song et al., 2019; Lee et al., 2018). Recent research has highlighted the remarkable robustness of fine-tuned language models to label noise. For example, Tänzer et al. (2022) find that pre-trained models such as BERT are more robust to noise. However, this generalization capacity comes at the cost of lower $F_1$ scores in the face of extreme class imbalances when no noise is present. Zhu et al. (2022) demonstrate that existing noise handling methods do not improve the peak performance of BERT models. Importantly, prior investigations primarily focus on assessing the impact of label noise on fully fine-tuned models within balanced datasets.

In this paper, our primary focus is on assessing whether LoRA tuning maintains robustness to noise inherent in the original model through fine-tuning. Additionally, we delve into the practical implications of both LoRA and fine-tuning methodologies by exploring scenarios involving imbalanced training data. Through comprehensive experimentation across datasets with varying noise levels and imbalances, our results demonstrate that LoRA tuning effectively preserves robustness against random label noise, matching the robustness observed in models subjected to full fine-tuning. This underscores LoRA's parameter efficiency comes without compromising model robustness. Notably, unlike full fine-tuning, which tends to overfit noisy samples along with clean ones, LoRA's training performance stabilizes at lower values as noise intensity increases. We meticulously monitor the influence of noisy and clean samples during training, revealing that LoRA predominantly learns from clean

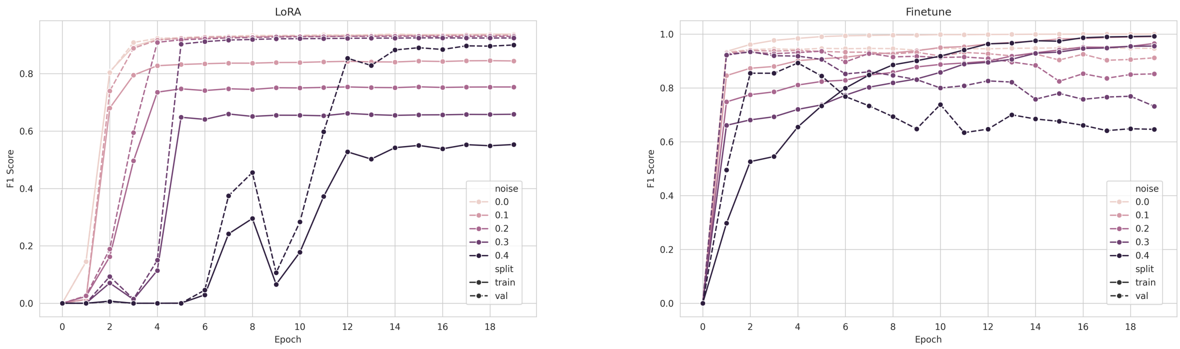



Figure 1: Comparison of learning dynamics for LoRA (left) and fine-tuning (right) on a balanced subset of the IMDB dataset. Both LoRA and fine-tuning exhibit robustness to noise, achieving high validation performances. However, LoRA demonstrates a distinctive resistance to overfitting the noise.

samples. Furthermore, our analysis of learning and forgetting events highlights LoRA's superior ability to retain learned information amidst increasing noise levels compared to full fine-tuning. We also scrutinize the model's resilience under substantial label imbalance and observe a marked decline in validation performance as data imbalance worsens, with this decline initiating at lower noise levels, particularly when the imbalance is more pronounced.

Overall, this study paves the way for understanding LoRA's potential in real-world scenarios with noise and imbalance. Our results demonstrate that LoRA tuning emerges as a robust and efficient contender for fine-tuning even in the presence of noisy labels. It retains the impressive noise resistance of its full-fine-tuning counterparts while showcasing unique advantages. Notably, LoRA learns primarily from clean data, exhibiting lower forgetting rates than fine-tuning under noise.

# 2 Background

## 2.1 Sources of Label Noise

Label noise is common in tasks involving human experts due to various factors ranging from insufficient evidence to perceptual errors (McNicol, 2005). Frénay and Verleysen (2013) categorize potential sources for label noise into four categories. Firstly, the information provided to annotators may lack sufficient detail, leading to unreliable labeling. For example, the annotation manual may not be elaborate or prescriptive enough (Rottger et al., 2022). Secondly, errors may also stem from non-experts often hired through crowdsourcing platforms to reduce annotation costs. Thirdly, many tasks, such as offensive language detection, are inherently subjective, where a single ground truth does not exist, leading to considerable variation in labels assigned by individual annotators. Lastly, label noise may occur due to data encoding issues (e.g., a post might be flagged as offensive because of accidental clicks)

## 2.2 Robustness to Noisy Labels

Deep learning approaches are known to suffer significant performance degradation when faced with noisy labels. This is because these approaches have the capacity to overfit an entire noisy training dataset, regardless of the level of noise present (Zhang et al., 2016, 2021). As a result, various methods have been proposed to mitigate the negative impact of noisy labels. These approaches can be broadly categorized into four categories; robust architectures, robust regularization, robust loss design, and sample selection (Song et al., 2022).

Limited research in NLP has investigated the susceptibility of models to the negative impacts of noisy labels. For instance, Jindal et al. (2019) show that CNN models used in text classification tend to overfit noisy labels, leading to a decrease in generalization performance. They demonstrated that adding a noise adaptation layer can significantly reduce the adverse effects of noisy labels. On the contrary, Transformers have exhibited remarkable resilience to noisy labels (Tänzer et al., 2022; Zhu et al., 2022). However, much of this research focuses on common benchmark NLP datasets with balanced label distributions, raising questions about whether this robustness persists in more practical settings with heavy label imbalance.

## 2.3 Parameter Efficient Tuning Methods

Methods for PEFT have become an important area of research in addressing the challenges stemming from the massive parameter size of large language models (Fu et al., 2023). PEFT methods involve maintaining the model parameters in a frozen state, and primarily operate by updating only a limited set of additional parameters within the model (He et al., 2022). These methods allow for rapid adaptation to new tasks without experiencing catastrophic forgetting (Pfeiffer et al., 2021) and frequently demonstrate enhanced robustness in out-of-distribution evaluation (Li and Liang, 2021).

Various approaches have been proposed for PEFT in recent years (Lester et al., 2021; Li and Liang, 2021; Hu et al., 2023, 2021). Out of these approaches, LoRA (Hu et al., 2022) has been one of the most widely adopted. LoRA is designed with the Lottery Ticket Hypothesis (LTH; Frankle and Carbin, 2018) in mind. According to the LTH, within densely connected, randomly initialized, feed-forward networks, there exist smaller subnetworks that, when trained independently, can achieve performance comparable to the original network. LoRA operationalizes LTH by approximating the model parameter updates with low-rank matrices inserted between every layer of Transformers. While these methods enable more efficient adaptation, investigating whether PEFT methods retain the capabilities and behaviors of the full model, especially in regard to robustness to noisy labels, will provide insights into the trade-offs between efficiency and model reliability.

# 3 Experimental Setup

We compare the performance of fine-tuning and LoRA-tuning of pre-trained language models when applied to training data that contain various degrees of noisy labels. To create datasets with varying levels of label noise, we randomly change the label of a data point with different probabilities ranging from 10% to 40%. This process, where the label corruption process is conditionally independent of the data, is known as instance-independent label noise (Song et al., 2022).

We conducted our experiments on the IMDB dataset (Maas et al., 2011), and limited the training data size to 10000 samples. For all experiments, we kept the evaluation and test sets fixed. We use the RoBERTa-base (Liu et al., 2019) and train all models for 20 epochs with a learning rate of 1e-5 and a linear scheduler of 0.06. We used AdamW optimizer (Loshchilov and Hutter, 2018) with an $L_2$ regularization of 0.01. For LoRA we used an $\alpha$ value of 16 and an $r$ value of 8.

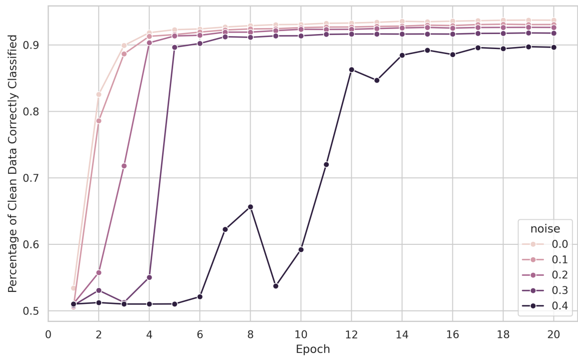



Figure 2: Percentage of clean samples correctly classified by LoRA. LoRA demonstrates a consistent ability to learn almost exclusively from the clean samples.

## 3.1 LoRA is Also Robust to Label Noise

First, we compare the train and validation performance of LoRA and fine-tuning on the fully balanced IMDB training dataset with various levels of label noise. Our goal in this analysis is to investigate whether LoRA exhibits similar patterns of robustness to full fine-tuning. As shown in Figure 1, similar to full fine-tuning, LoRA achieves high validation performance of above 90% regardless of the level of noise present. However, the two methods behave differently on the training data. Specifically, we observe that full fine-tuning overfits all training data (including the noisy samples) consistently getting $F_1$ scores of above 95% on the noisy training set. However, the training performance of LoRA plateaus. Furthermore, we observe that the maximum training performance of LoRA decreases from 93.8% to 55.3% as we increase the noise in the training dataset (see table Table 1 for detailed results). This low performance on the noisy training set, in addition to high validation performance, suggests that LoRA might only be learning to predict the clean samples correctly.

To gain deeper insights into the underlying mechanisms leading to LoRA's robustness, we look into the accuracy of the model over both the noisy and clean sets as training progresses. Figure 2 shows what percentage of correctly classified samples are clean data points during the training. We observe that as training progresses, over 90% of correctly classified data points come from the clean set. However, a stark contrast emerges when considering its performance with noisy samples. Despite the varying levels of noise, the model consistently resists fitting the noisy data, accurately classifying as few as 10% of the noisy samples (Figure 9).

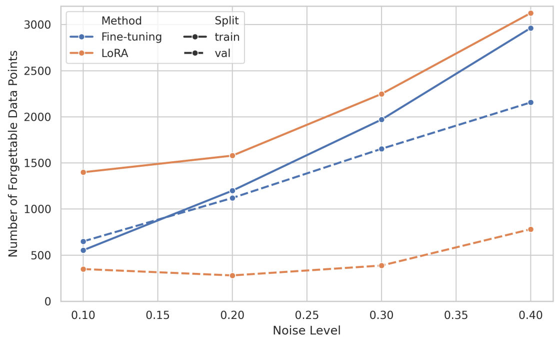



Figure 3: Number of forgettable data points for LoRA (blue) and fine-tuning (orange). LoRA consistently forgets fewer data points on the validation set.

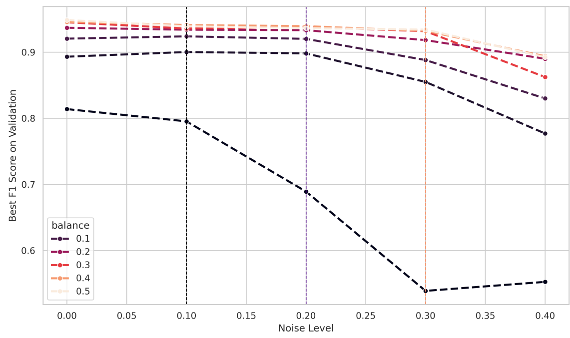



Figure 4: The best validation performance degradation happens for lower values of noise as imbalance worsens

### 3.2 Learning and Forgetting in LoRA

The total number of forgettable datapoints reveals how models get impacted from noise over training, and points to their resilience to noisy labels (i.e., a model that forgets fewer datapoints as a result of increased noise can potentially generalize better even after facing noisy examples). Here, we define *forgettable* data points for a model as those initially learned during training (i.e., correctly classified at some point), yet subsequently forgotten (i.e., misclassified in the learning process). Figure 3 shows the number of forgettable data points for LoRA and fine-tuning for various levels of noise. Notably, LoRA consistently exhibits a low number of forgettable data points on the validation set, indicating its robustness, whereas the number of forgettable data points increases for fine-tuning as the level of noise over training data worsens. Both models exhibit similar trends for forgettable data points on the noisy training data, with the count increasing as the noise level rises.

### 3.3 Robustness in the Face of Data Imbalance

Many real-world NLP applications lack balanced data distributions. For example, datasets for hate speech or offensive language detection often have a small fraction of positive samples (Yin and Zubiaga, 2021). To better understand the benefits of the observed robustness to label noise in practical settings, it's crucial to acknowledge the prevalence of imbalanced data. To assess this, we constructed various versions of the IMDB dataset, keeping the training size constant at 10000 but varying the percentage of positive sentiment samples between 50%, 40%, 30%, 20%, 10%, and 5%. For each version of the imbalanced dataset, we added varying degrees of noise conducted robustness to noise experiments as described in section 3.

As depicted in Figure 4, compared to validation performance with no noise, the validation performance drops more as the imbalance intensifies. For example, the performance degrades by 5.2% when 40% of noise is added to the balanced dataset. However, this degradation is intensified to 12% with the same noise when the dataset is balanced at 5%. This widening gap underscores the challenge posed by imbalanced data and emphasizes the importance of developing robust NLP models capable of handling such scenarios effectively. Furthermore, we observe that this performance gap begins to manifest even at lower levels of noise in the data distribution. This early emergence of performance discrepancies highlights the sensitivity of NLP models to imbalanced datasets, suggesting that even a modest degree of imbalance can significantly impact model generalization.

## 4 Conclusion

Our study highlights the efficacy and resilience of PEFT, particularly LoRA, in learning from noisy labels. Through our comprehensive analysis, we have shown that LoRA tuning not only retains the robustness to label noise exhibited by fine-tuning but also demonstrates unique advantages. Specifically, LoRA shows resistance to overfitting noisy labels, an ability to learn almost exclusively from clean data, and lower forgetting rates compared to fine-tuning. Additionally, our experiments shed light on label noise robustness in imbalanced training data. We found that imbalanced data exacerbates the effects of noisy label, particularly as the level of imbalance increases, even at lower noise levels. These findings highlight LoRA's potential in real-world scenarios where noisy data and class imbalances prevail, offering a promising balance between efficiency and robustness for adapting large-scale language models to downstream tasks.

## 5 Limitation

Our analysis is limited to English. Hence, the conclusions drawn may not fully translate to other languages or linguistic contexts due to differences in syntax, semantics, among other factors. Consequently, the applicability of our findings in multilingual or cross-cultural settings warrants careful consideration and potentially necessitates additional research to ascertain their broader relevance. Additionally, we acknowledge that the IMDB dataset is not devoid of noisy labels. However, since this dataset has been widely adopted in machine learning research, the extent of noise can be assumed to be limited. We also acknowledge that our analysis is limited in the type of noise explored. Variations in the nature of noise, such as instance-dependent noise could lead to disparate results not explored within the scope of this work. We believe that our analysis and experimental design serve as a solid foundation for future researchers to explore other noise structures, such as instance-dependent noise. In summary, while our study provides valuable insights within the confines of our chosen language models, methods, datasets, noise types, and linguistic context, it is essential to recognize the limitations inherent in these choices. Future research endeavors should aim to address these limitations by exploring alternative approaches, diverse datasets, and broader linguistic contexts to enrich our understanding and enhance the generalizability of our findings

## References


Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jonathan Frankle and Michael Carbin. 2018. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*.

Benoît Frénay and Michel Verleysen. 2013. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869.

Zihao Fu, Haoran Yang, Anthony Man-Cho So, Wai Lam, Lidong Bing, and Nigel Collier. 2023. On the effectiveness of parameter-efficient fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12799–12807.

Gemini GeminiTeam, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2021. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Lee. 2023. LLM-adapters: An adapter family for parameter-efficient fine-tuning of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5276, Singapore. Association for Computational Linguistics.

Ishan Jindal, Daniel Pressel, Brian Lester, and Matthew Nokleby. 2019. An effective label noise model for DNN text classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3246–3256, Minneapolis, Minnesota. Association for Computational Linguistics.

Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. 2018. Cleannet: Transfer learning for scalable image classifier training with label noise. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5447–5456.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In

*Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Don McNicol. 2005. *A primer of signal detection theory*. Psychology Press.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Hwanjun Song, Minseok Kim, and Jae-Gil Lee. 2019. Selfie: Refurbishing unclean samples for robust deep learning. In *International Conference on Machine Learning*, pages 5907–5915. PMLR.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2022. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Michael Tänzer, Sebastian Ruder, and Marek Rei. 2022. Memorisation versus generalisation in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7564–7578, Dublin, Ireland. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2021. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115.

Dawei Zhu, Michael A. Hedderich, Fangzhou Zhai, David Adelani, and Dietrich Klakow. 2022. Is BERT robust to label noise? a study on learning with noisy labels in text classification. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 62–67, Dublin, Ireland. Association for Computational Linguistics.

## A Hardware

All the experiments were conducted on an NVIDIA RTX A6000 with 48GB RAM. Each epoch takes around 10 minutes to run on a single GPU.

## B Detailed Results for Robustness to Noise

| | **LoRA $F_1$** | | **Fine-Tuning $F_1$** | |
|---|---|---|---|---|
| **Noise** | **Train** | **Val** | **Train** | **Val** |
| 0% | 0.938 | 0.938 | 1.00 | 0.949 |
| 10% | 0.845 | 0.934 | 0.991 | 0.939 |
| 20% | 0.754 | 0.931 | 0.965 | 0.936 |
| 30% | 0.662 | 0.925 | 0.955 | 0.934 |
| 40% | 0.553 | 0.900 | 0.992 | 0.893 |

Table 1: $F_1$ scores of LoRA and fine-tuning on balanced IMDB dataset for various degrees of noise.

## C LoRA Almost Exclusively Learns from the Clean Data

Figure 9 illustrates the accuracy comparison between LoRA and fine-tuning on the noisy samples of the training set. A notable observation is the strikingly opposite patterns exhibited by the two approaches. LoRA consistently yields a lower accuracy, typically less than 10%, on the training data. Conversely, fine-tuning demonstrates the capability

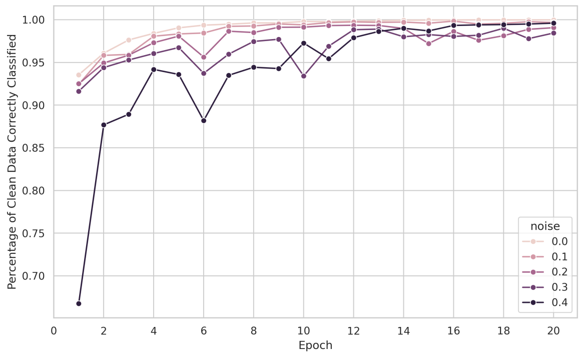



Figure 5: Percentage of clean samples correctly classified by fine-tuning.

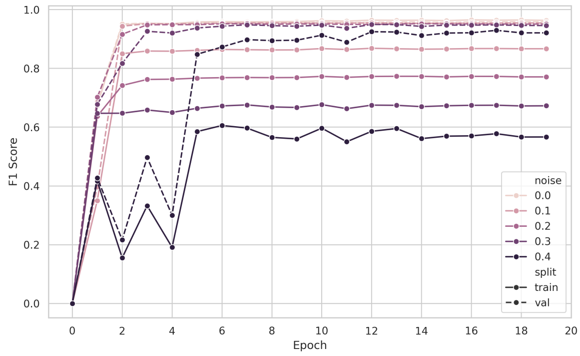



Figure 6: Learning dynamics for LoRA applied to RoBERTa-large on a balanced subset of the IMDB.

to adapt to noisy data irrespective of the noise level, achieving an accuracy of approximately 90% on both the noisy and clean subsets (Figure 5).

## D Learning and Forgetting

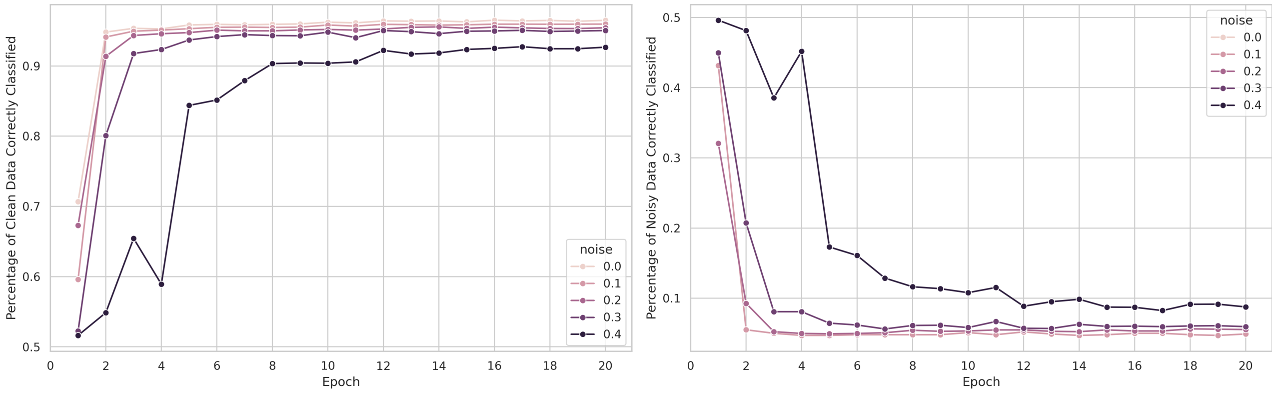
In addition to performance, we track when data points are correctly classified for the first time (*learning event*) and when a data point that was previously learned is misclassified by the model (*forgetting event*). Figure 10 presents a comparison of learning events in LoRA and fine-tuning. It is evident from the graph that in both approaches, the majority of learning events occur during the initial epoch, with LoRA consistently having fewer learning events compared to fine-tuning in these early stages. Yet, as shown in the figure, LoRA exhibits more learning events in later epochs compared to fine-tuning, especially in scenarios with higher noise levels. Figure 11 provides a comparison of forgetting events in LoRA and fine-tuning. We observe a clear distinction between the two approaches; namely, fine-tuning shows higher forgetting events throughout the training, especially for higher values of noise compared to LoRA.

## E Increasing Model Size

To examine the influence of model size on robustness, we additionally conduct the analysis outlined in section 3 using RoBERTa-large. Looking at Figure 6 we observe similar patterns of robustness to noise to RoBERTA-base, the only notable difference is that RoBERTa-large plateaus at earlier epochs compared to RoBERTa-base.

As depicted in Figure 7, the accuracy of RoBERTa-large on both clean and noisy training subsets is shown for different levels of noise. We note a pattern similar to RoBERTa-base.

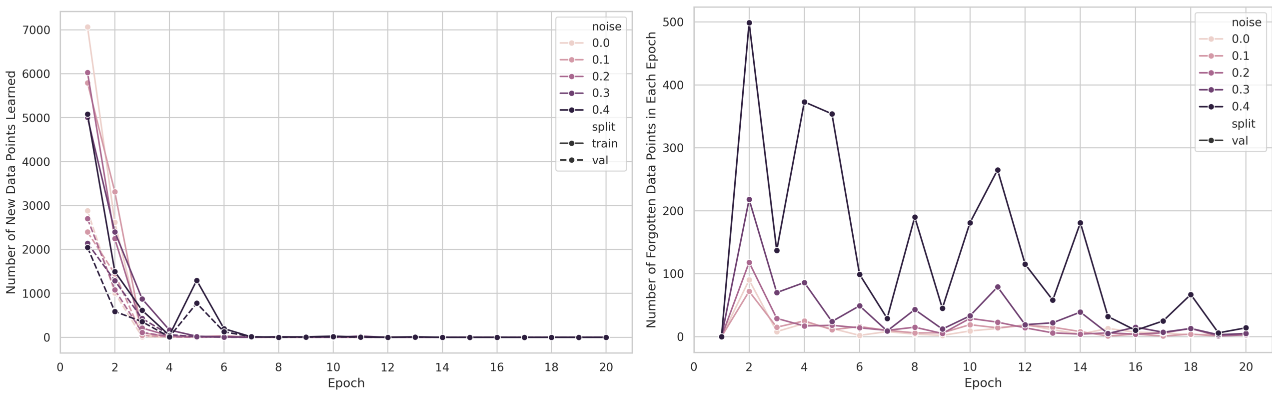
As shown in Figure 8, LoRA-tuning RoBERTa-large also exhibits notable ability in fitting clean samples while demonstrating resilience against overfitting noisy samples. However, we observe that the larger model learns the clean data (and unlearns noisy data) at earlier epochs compared to the base model.




(a) Clean

(b) Noisy

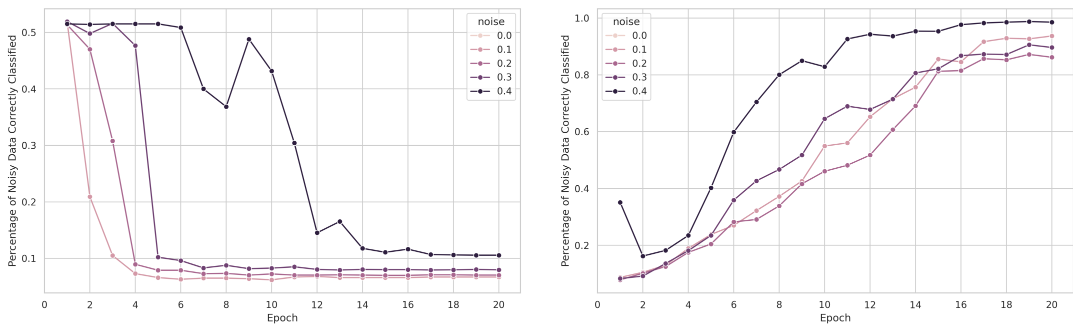
Figure 7: Comparison of the accuracy on clean (left) and noisy (right) samples in the training set for LoRA applied to RoBERTa-large on balanced IMDB dataset.




(a) Learning

(b) Forgetting

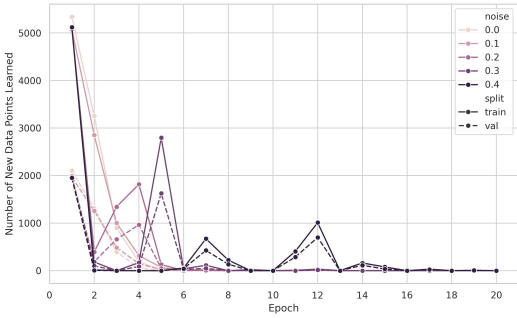
Figure 8: Comparison of the accuracy on learning (right) and forgetting (left) for LoRA applied to RoBERTa-large on balanced IMDB dataset.



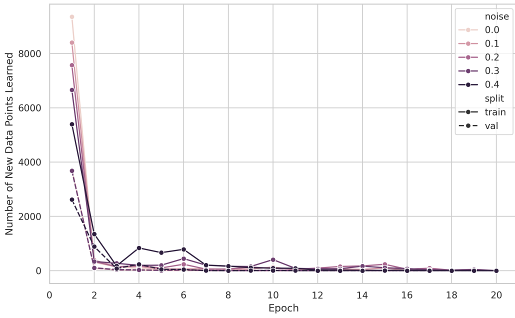

(a) LoRA

(b) Fine-tuning

Figure 9: Comparison of the accuracy on noisy samples in the training set for LoRA (left) and fine-tuning (right) on balanced IMDB dataset.

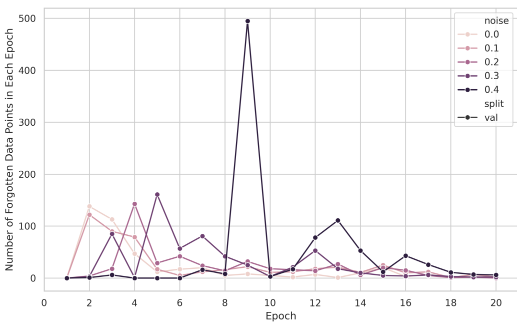



(a) LoRA

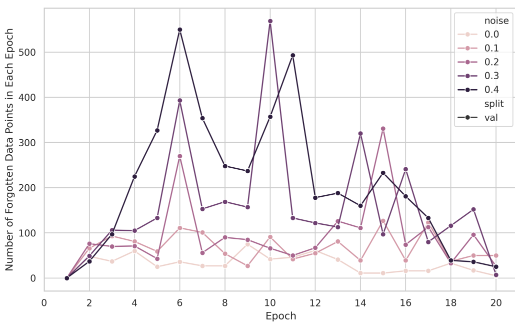



(b) Fine-tuning

Figure 10: Comparison of learning events for LoRA (left) and fine-tuning (right) on balanced IMDB dataset.




(a) LoRA




(b) Fine-tuning

Figure 11: Comparison of forgetting events for LoRA (left) and fine-tuning (right) on balanced IMDB dataset.