DSA4213 Assignment 3 Report

Fong Kang Wei

16 October 2025

1 Introduction

Transformer-based language models such as BERT and DistilBERT have become foundational in modern NLP, achieving strong results across text classification and question answering tasks. However, fully fine-tuning all model parameters is computationally demanding and memory-intensive. Parameter-efficient methods like Low-Rank Adaptation (LoRA) address this limitation by updating only a small subset of trainable parameters, thereby reducing memory usage while maintaining competitive accuracy.

This project compares **Full Fine-Tuning** and **LoRA Fine-Tuning** on the IMDb movie reviews dataset for binary sentiment classification. Both approaches are evaluated using *accuracy* and *F1 score* to assess predictive performance, while *training time* and *peak memory* measure computational efficiency. Qualitative evaluation further analyzes misclassified examples and reflects on implementation insights.

The objective is to understand the trade-offs between performance and efficiency, identify patterns in model behavior, and highlight practical takeaways and limitations when deploying transformer models under constrained computational resources. To ensure reproducibility, a fixed random seed was used for all runs. All code is provided in the accompanying `dsa4213_assignment3.ipynb`, and the complete repository is available at `https://github.com/kangwei01/fongkangwei-dsa4213-assignment3.git`.

2 Data

2.1 Dataset Description

The dataset used in this study is the **IMDb movie reviews dataset**, obtained from the `datasets` library provided by Hugging Face using `load_dataset("imdb")`. A widely adopted benchmark for sentiment classification, it contains 50,000 English-language movie reviews, evenly split into 25,000 training and 25,000 testing samples, with binary sentiment labels of positive (1) and negative (0).

Motivation. The IMDb dataset was selected for three main reasons. First, it provides a clear and interpretable benchmark for evaluating model performance on text classification. Second, its balanced label distribution allows for fair comparison of evaluation metrics such as accuracy and F1 score without bias toward a particular class. Finally, the dataset's moderate size makes it suitable for experimentation on limited hardware resources, enabling efficient testing of both full fine-tuning and parameter-efficient fine-tuning methods like LoRA on a local machine.

2.2 Preprocessing

The IMDb dataset was preprocessed using the Hugging Face `datasets` and `transformers` libraries. A preview of the dataset is shown below.

```
Example 1
Text: With these people faking so many shots, using old footage, and gassing animals to get them out, not to m
ention that some of the scenes were filmed on a created set with actors, what's to believe? Old film of countr
ies is nice, but the animal abuse and degradation of natives is painful to watch in th ...
Label: negative

Example 2
Text: I don't know the stars, or modern Chinese teenage music - but I do know a thoroughly entertaining movie
when I see one.<br /><br />Kung Fu Dunk is pure Hollywood in its values - it's played for laughs, for love, an
d is a great blend of Kung Fu and basketball.<br /><br />Everybody looks like they had ...
Label: positive

Example 3
Text: Having lived in Ontario my whole life, in the same town that Marlene Moore grew up in, I've heard storie
s of her from my parents, grandparents and family members. So when I found out that they would be filming a mo
vie about her, and that the beginning would be shot on my street, and her house quite  ...
Label: positive
```

Figure 1: Three classification examples from the IMDb dataset

Splits. A validation split comprising 10% of the original training data was created to support model selection and hyperparameter tuning. For local computation efficiency, smaller subsets of the training, validation, and test splits were randomly selected while maintaining balanced label distributions.

Tokenization. Given that IMDb reviews often contain lengthy narratives, the input sequences were truncated to a maximum length of 256 tokens. This provides a balance between preserving sufficient textual context and maintaining computational efficiency during fine-tuning. Padding to the maximum sequence length was applied to stabilize batch sizes and accelerate computation on the Apple MPS backend. Texts were tokenized using the `DistilBERT` tokenizer in uncased mode to ensure consistency in handling capitalization.

Model Choice. `DistilBERT` was chosen as the base model because it offers a strong balance between performance and resource efficiency. As a distilled version of BERT, it retains most of BERT's language understanding capability while being much smaller and faster to fine-tune. This makes it ideal for experimentation on limited hardware without sacrificing downstream accuracy.

3 Hyperparameter Tuning

3.1 Full Fine-Tuning

Hyperparameter tuning for full fine-tuning was conducted with a fixed-budget grid search over three key factors: learning rate, batch size, and training epochs. Concretely, we evaluated `learning_rate` $\in \{2\times10^{-5}, 3\times10^{-5}, 5\times10^{-5}\}$, `per_device_train_batch_size` $\in \{8, 16\}$, and `num_train_epochs` $\in \{2, 3\}$, yielding 12 trials. Each trial used the same tokenization settings and validation split.

After training, the model was evaluated on the validation set and we recorded *(i)* validation F1 and accuracy, and *(ii)* efficiency signals measured via an in-run profiler: wall-clock time per trial (`time_s`) and peak host memory (`peak_mem_gb`). Results were then ranked by validation F1; ties were broken by shorter wall-clock time (and then lower validation loss). Table 1 provides the results, with the best-performing configuration highlighted in bold.

eval_loss	eval_accuracy	eval_f1	learning_rate	batch_size	num_epochs	time_s	peak_mem_gb
0.284	**0.899**	**0.899**	**2e-5**	**16**	**2**	**795.7**	**1.08**
0.281	0.894	0.894	3e-5	16	2	676.5	0.77
0.334	0.892	0.892	5e-5	16	2	624.4	0.51
0.350	0.892	0.892	2e-5	16	3	1899.4	0.67
0.467	0.890	0.890	2e-5	8	3	1172.2	0.90
0.447	0.890	0.890	5e-5	16	3	972.2	0.47
0.362	0.889	0.889	2e-5	8	2	816.7	0.95
0.387	0.888	0.888	3e-5	8	2	980.2	0.69
0.406	0.886	0.886	3e-5	16	3	973.9	0.69
0.540	0.885	0.885	5e-5	8	3	1091.4	0.49
0.412	0.881	0.881	5e-5	8	2	662.6	0.51
0.535	0.881	0.881	3e-5	8	3	1127.6	0.75

Table 1: Validation results for Full Fine-Tuning hyperparameter grid search

3.2 LoRA Fine-Tuning

Instead of updating all model weights, LoRA injects low-rank matrices into the attention layers and optimizes only these additional parameters while keeping the base model frozen, thereby significantly reducing the number of trainable parameters.

We used the `peft` library to configure LoRA with rank $r = 8$, scaling factor $\alpha = 16$, and dropout $p = 0.1$, targeting the query and value projection layers (`["q_lin", "v_lin"]`). These settings follow common practice in literature and were found to balance stability and efficiency on small-scale devices.

Hyperparameter tuning followed the same grid search procedure as in full fine-tuning, varying learning rate $\{1 \times 10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}\}$, batch size $\{8, 16\}$, and number of epochs $\{2, 3\}$. Each configuration was trained and evaluated on the validation set, recording F1, accuracy, total training time (`time_s`), and peak memory usage (`peak_mem_gb`).

By freezing the backbone weights and updating only the injected LoRA layers, the training process achieved notably lower memory consumption and faster iteration times on average (Table 2). Again, the best configuration was selected based on validation F1, with ties broken by shorter wall-clock time.

eval_loss	eval_accuracy	eval_f1	learning_rate	batch_size	num_epochs	time_s	peak_mem_gb
0.2795	**0.898**	**0.8980**	**5e-4**	**16**	**3**	**1184.59**	**0.49**
0.2794	0.892	0.8920	2e-4	16	3	929.19	0.46
0.2726	0.889	0.8890	5e-4	16	2	647.68	0.48
0.3327	0.888	0.8880	5e-4	8	3	951.62	0.50
0.2901	0.887	0.8870	5e-4	8	2	629.56	0.44
0.3153	0.887	0.8870	2e-4	8	3	913.21	0.48
0.2876	0.886	0.8860	1e-4	16	3	765.35	0.47
0.2854	0.885	0.8850	2e-4	16	2	593.85	0.50
0.2961	0.885	0.8850	2e-4	8	2	542.63	0.49
0.3020	0.883	0.8830	1e-4	8	3	812.91	0.46
0.3016	0.882	0.8820	1e-4	8	2	557.50	0.49
0.3061	0.876	0.8760	1e-4	16	2	558.33	0.47

Table 2: Validation results for LoRA Fine-Tuning hyperparameter grid search

4 Model Test Evaluation & Comparisons

Using the best hyperparameter configurations identified from the validation tuning phase (`learning_rate=2e-5`, `batch_size=16`, `num_epochs=2` for Full Fine-Tuning, and `learning_rate=5e-4`, `batch_size=16`, `num_epochs=3` for LoRA), both strategies were retrained from the same pretrained `DistilBERT-base-uncased` checkpoint and evaluated on the held-out IMDb test set. To ensure consistency, the same preprocessing pipeline, tokenization, and random seed were used across runs. During training, efficiency profiling recorded total wall-clock time and peak host memory consumption, providing quantitative measures for computational efficiency alongside test accuracy and F1 score. Table 3 shows the summary results for both strategies.

Table 3: Summary test results for Full Fine-Tuning and LoRA

Strategy	Test loss	Test acc.	Test F1	Wall-clock time (s)	Avg. epoch time (s)	Peak mem. (GB)
Full	0.312	0.874	0.874	809.21	404.61	0.57
LoRA	0.315	0.883	0.883	850.07	283.36	0.46

4.1 Accuracy & F1

Figure 2 summarizes the test performance of both fine-tuning strategies on the IMDb dataset. Surprisingly, the LoRA model achieved slightly higher accuracy and F1 scores (0.883) compared to Full Fine-Tuning (0.874), despite updating only a small subset of parameters.

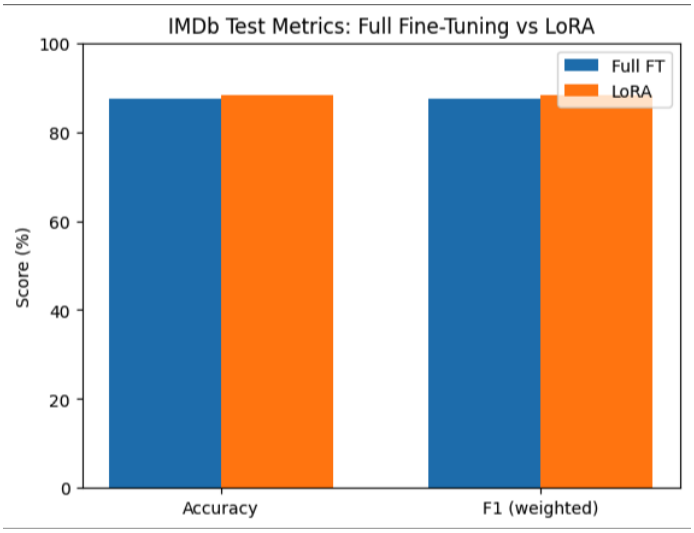


Figure 2: Test accuracy and F1 comparison between Full Fine-Tuning and LoRA

Both models reported identical accuracy and F1 scores within their respective runs. This is expected in **balanced** binary classification tasks where the dataset has an even distribution of positive and negative reviews and the model predictions are well-calibrated across both classes. Under these conditions, precision and recall contribute equally, resulting in the weighted F1 score aligning closely with accuracy.

These findings suggest that LoRA not only preserves the representational strength of the pretrained model but may also benefit from implicit regularization effects that mitigate overfitting. Overall, the results demonstrate that parameter-efficient fine-tuning can match—and even slightly outperform—full fine-tuning, offering a compelling trade-off between performance and computational efficiency.

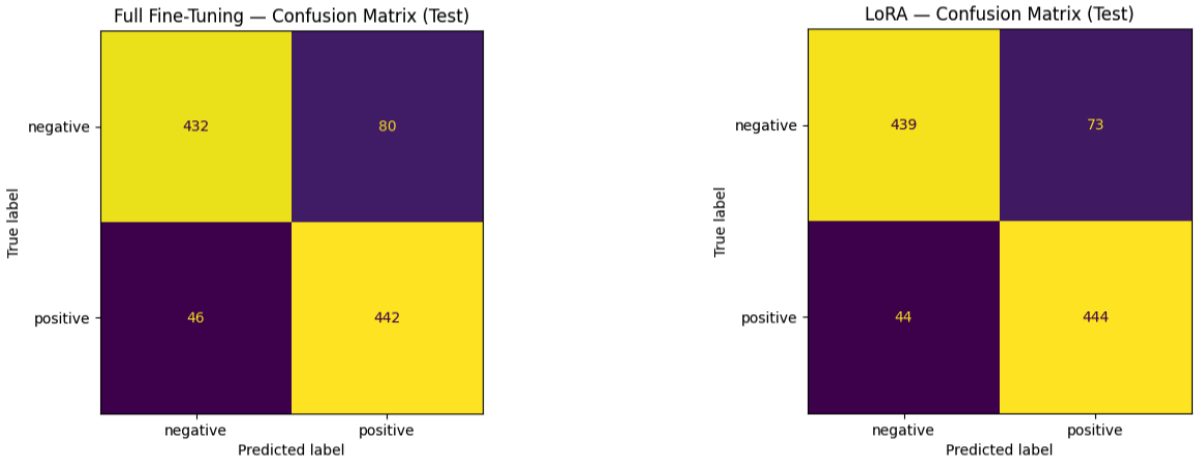
4.2 Classification Reports & Confusion Matrices

Table 4 presents the precision, recall, and F1-scores for both sentiment classes. LoRA achieved slightly higher scores across all metrics.

Model / Class	Precision	Recall	F1-Score	Support	Accuracy
Full Fine-Tuning					
Negative	0.9038	0.8438	0.8727	512	
Positive	0.8467	0.9057	0.8752	488	
Macro Avg	0.8753	0.8747	0.8740	1000	
Weighted Avg	0.8759	0.8740	0.8740	1000	**0.8740**
LoRA Fine-Tuning					
Negative	0.9089	0.8574	0.8824	512	
Positive	0.8588	0.9098	0.8836	488	
Macro Avg	0.8839	0.8836	0.8830	1000	
Weighted Avg	0.8845	0.8830	0.8830	1000	**0.8830**

Table 4: Classification report comparison between Full Fine-Tuning and LoRA

Figure 3 illustrates the class-wise prediction breakdown for both models. LoRA produced a slightly more balanced confusion matrix, with **7 fewer false positives** and **2 fewer false negatives** compared to Full Fine-Tuning. This indicates that LoRA made marginally fewer misclassifications for both classes while preserving strong overall accuracy. These results align with the higher precision and recall observed in Table 4, suggesting that LoRA generalizes better despite its reduced parameter footprint.



(a) Full Fine-Tuning

(b) LoRA Fine-Tuning

Figure 3: Confusion matrices comparing Full Fine-Tuning and LoRA

4.3 Error Cases

To better understand model behavior, we qualitatively examined ten misclassified samples from each fine-tuning strategy. The majority of errors could be attributed to three main causes: **(1)** mixed sentiment

reviews, **(2)** sarcastic language, and **(3)** context truncation due to token length limits. Representative examples for each category are provided below.

1. Mixed Sentiment Reviews. Reviews that contained both positive and negative statements often confused the models, as sentiment cues appeared in opposing directions.

True Label	**Predicted Label**	**Truncated Review (first 256 tokens)**
Negative	Positive	*"Lovely music. Beautiful photography, some of scenes are breathtaking and affecting. But the dramatic tension is lost in a film that is so poorly edited it is hard to know what exactly is going on. At times, the dialogue is incomprehensible. Then there is Richard Gere. He's supposed to be a factory worker who gets into trouble and gets work on a farm. We see dozens of farmhands sweaty and dirty in ..."*
Negative	Positive	*"Intended as light entertainment, this film is indeed successful as such during its first half, but then succumbs to a rapidly foundering script that drops it down. Harry (Judd Nelson), a "reformed" burglar, and Daphne (Gina Gershon), an aspiring actress, are employed as live window mannequins at a department store where one evening they are late in leaving and are locked within, whereupon they wit ..."*
Positive	Negative	*"i was having a horrid day but this movie grabbed me, and i couldn't put it down until the end... and i had forgotten about my horrid day. and the ending... by the way... where is the sequel!!!¡br /¿¡br /¿the budget is obviously extremely low... but ... look what they did with it! it reminds me of a play... they are basically working with a tent, a 'escape pod', a few guns, uniforms, camping gear, ..." (In this case, 'horrid' was in reference to the reviewer's day, unrelated to the movie.)*

Table 5: Examples of mixed-sentiment reviews that led to misclassifications

2. Sarcastic Reviews. The models struggled with sarcasm and irony, where literal word meanings contradicted the intended sentiment. This limitation reflects the difficulty of detecting pragmatic tone without explicit contextual cues.

True Label	**Predicted Label**	**Truncated Review (first 256 tokens)**
Negative	Positive	*"A truly masterful piece of filmmaking. It managed to put me to sleep and to boggle my mind. So boring that it induces sleep and yet so ludicrous that it made me wonder how stuff like this gets made. Avoid at all costs. That is, unless you like taking invisible cranial punishment, in which case I highly recommend it. ..."*
Negative	Positive	*"This is, without a doubt, the most hilarious movie I've ever seen. Seriously, if the makers of this movie are ever discovered, they'll put guys like Jim Carrey out of a job. Rent "Jack-O" tonight! Believe me, you won't regret it! ..."*

Table 6: Examples of sarcastic reviews misinterpreted as positive sentiment

3. Context-Length Truncation. Due to computational limits, reviews were truncated to a maximum of 256 tokens. Some misclassifications arose because key sentiment-bearing phrases occurred beyond this limit or were buried after lengthy neutral plot summaries.

True Label	Predicted Label	Truncated Review (first 256 tokens)
Negative	Positive	*“”An astronaut (Michael Emmet) dies while returning from a mission and his body is recovered by the military. The base where the dead astronaut is taken to becomes the scene of a bizarre invasion plan from outer space. Alien embryos inside the dead astronaut resurrect the corpse and begin a terrifying assault on the military staff in the hopes of conquering the world,” according to the DVD sleeve’s …”*
Negative	Positive	*“This is about some vampires (who can run around out in the sunlight), that are causing some problems down in South America. Casper Van Dien is sent in with his team of commandos to investigate. The movie opens with Van Dien & Co. walking through the jungle, and there’s this huge black guy who just absolutely, positively cannot act. He speaks all his lines as if he’s reading them off the cue-cards …”*

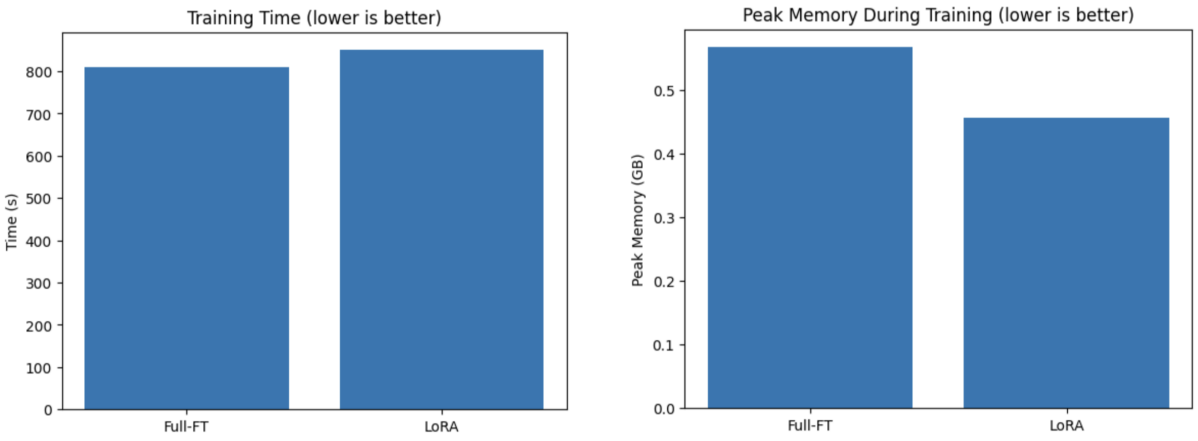
Table 7: Examples of reviews where truncation likely caused misclassification

Interestingly, both models shared most of the same misclassified samples, indicating that they struggled with similar linguistic complexities rather than architecture-specific weaknesses. All ten misclassified samples for both Full Fine-Tuning and LoRA are provided in the Appendix for reference.

4.4 Efficiency

To assess computational efficiency, both strategies were compared in terms of total wall-clock training time and peak memory usage. As shown in Table 3, LoRA recorded a longer overall training time (850 s) than Full Fine-Tuning (809 s). However, this difference is attributed to LoRA’s optimal configuration running for three epochs, compared to two for Full Fine-Tuning. When normalized per epoch, LoRA’s training time was substantially shorter (283 s vs. 405 s), demonstrating greater efficiency per iteration.

In terms of memory footprint, LoRA consumed less peak memory (0.46 GB) than Full Fine-Tuning (0.57 GB), consistent with its parameter-efficient design. This reduction in memory usage makes LoRA particularly appealing for low-resource environments such as personal laptops.



(a) Training time

(b) Peak memory consumption during training

Figure 4: Efficiency comparison between Full Fine-Tuning and LoRA

Overall, these findings suggest that LoRA offers a favorable trade-off between training efficiency and model performance—achieving higher accuracy and F1 than Full Fine-Tuning, while using less memory and maintaining shorter per-epoch runtime. In practical scenarios, this translates to slightly longer total training but greater cost-effectiveness and scalability.

5 Conclusion

5.1 Key Takeaways

The results demonstrate that parameter-efficient fine-tuning (PEFT) via LoRA can achieve performance on par with, and slightly exceeding, full fine-tuning while using less memory and fewer trainable parameters. However, both models showed similar error patterns that stemmed primarily from the inherent ambiguity of natural language and dataset characteristics, rather than from differences between full and PEFT strategies. Overall, LoRA provides a more resource-efficient alternative that delivers competitive accuracy, making it the preferable choice for practical deployment in constrained environments.

5.2 Limitations

The main limitation of this study stems from the local computational constraints. To fit within available hardware resources, input sequences were truncated—potentially omitting contextual cues critical for sentiment inference. Additionally, only a subset of the IMDb dataset was used, which may limit the statistical robustness of results.

Future experiments could explore longer context windows using more capable hardware, evaluate LoRA under different low-rank settings or hybrid PEFT methods, and test across larger or multi-domain datasets to validate the generality of these findings. Extending the analysis to multi-class or multi-label sentiment tasks could also provide deeper insights into the scalability of PEFT approaches.

AI Assistance Declaration

I acknowledge the use of ChatGPT-5 as an editing assistant in the preparation of this report. Specifically, ChatGPT was used to help outline the report structure, generate boilerplate LaTeX (section headers, figure/table placeholders), and improve expression and phrasing for clarity and conciseness. All code implementation, plots, analyses, and interpretations were independently conducted and verified by me.

Appendix

Full Fine-Tuning Misclassifications

True Label	Predicted Label	Truncated Review (first 256 tokens)
Negative	Positive	*"Coming from Kiarostami, this art-house visual and sound exposition is a surprise. For a director known for his narratives and keen observation of humans, especially children, this excursion into minimalist cinematography begs for questions: Why did he do it? Was it to keep him busy during a vacation at the shore? ¡br /¿¡br /¿"Five, 5 Long Takes" consists of, you guessed it, five long takes. They a ..."*
Negative	Positive	*"Intended as light entertainment, this film is indeed successful as such during its first half, but then succumbs to a rapidly foundering script that drops it down. Harry (Judd Nelson), a "reformed" burglar, and Daphne (Gina Gershon), an aspiring actress, are employed as live window mannequins at a department store where one evening they are late in leaving and are locked within, whereupon they wit ..."*
Negative	Positive	*"An astronaut (Michael Emmet) dies while returning from a mission and his body is recovered by the military. The base where the dead astronaut is taken to becomes the scene of a bizarre invasion plan from outer space. Alien embryos inside the dead astronaut resurrect the corpse and begin a terrifying assault on the military staff in the hopes of conquering the world," according to the DVD sleeve's ...*
Negative	Positive	*Lovely music. Beautiful photography, some of scenes are breathtaking and affecting. But the dramatic tension is lost in a film that is so poorly edited it is hard to know what exactly is going on. At times, the dialogue is incomprehensible. Then there is Richard Gere. He's supposed to be a factory worker who gets into trouble and gets work on a farm. We see dozens of farmhands sweaty and dirty in ...*
Negative	Positive	*This is about some vampires (who can run around out in the sunlight), that are causing some problems down in South America. Casper Van Dien is sent in with his team of commandos to investigate. The movie opens with Van Dien & Co. walking through the jungle, and there's this huge black guy who just absolutely, positively cannot act. He speaks all his lines as if he's reading them off the cue-cards ...*
Negative	Positive	*A not so good action thriller because it unsuccessfully trends the same water as early Steven Seagal films because there is not a very good set piece. Steven Seagal plays the same kind of character that he has played since Above the Law. In my opinion the performance of Keenen Ivory Wayans is wasted in such an average film and belongs in a much better film. Bob Gunton is okay as the main heavy. Th ...*
Positive	Negative	*I remember seeing this film in the mid 80's thought it a well paced and well acted piece. I now work quite often in Berkeley Square and the had to get a copy of DVD to remind myself how little the area has changed, although my office is newish it just 30 seconds away from "the bank". Even Jack Barclays car dealership is still there selling Bentleys and Rolls Royces.¡br /¿¡br /¿It's look like the D ...*
Positive	Negative	*Sex, drugs, racism and of course you ABC's. What more could you want in a kid's show!¡br /¿¡br /¿————————————————————————————————¡br /¿¡br /¿"User Comment Guidelines ¡br /¿¡br /¿Please note there is a 1,000 word limit on comments. The recommended length is 200 to 500 words. The minimum length for comments is 10 lines of text. Comments which ar ...*
Negative	Positive	*I used to always love the bill because of its great script and characters, but lately i feel as though it has turned into an emotional type of soap. If you look at promotional pictures/posters of the bill now you will see either two of the officers hugging/kissing or something to do with friendships whereas promotional pictures of the bill a long time ago would have shown something to do with crim ...*
Negative	Positive	*A truly masterful piece of filmmaking. It managed to put me to sleep and to boggle my mind. So boring that it induces sleep and yet so ludicrous that it made me wonder how stuff like this gets made. Avoid at all costs. That is, unless you like taking invisible cranial punishment, in which case I highly recommend it. ...*

Table 8: Ten misclassified IMDb samples under Full Fine-Tuning.

LoRA Fine-Tuning Misclassifications

True Label	Predicted Label	Truncated Review (first 256 tokens)
Negative	Positive	*Coming from Kiarostami, this art-house visual and sound exposition is a surprise. For a director known for his narratives and keen observation of humans, especially children, this excursion into minimalist cinematography begs for questions: Why did he do it? Was it to keep him busy during a vacation at the shore? ¡br /¿¡br /¿"Five, 5 Long Takes" consists of, you guessed it, five long takes. They a ...*
Negative	Positive	*Intended as light entertainment, this film is indeed successful as such during its first half, but then succumbs to a rapidly foundering script that drops it down. Harry (Judd Nelson), a "reformed" burglar, and Daphne (Gina Gershon), an aspiring actress, are employed as live window mannequins at a department store where one evening they are late in leaving and are locked within, whereupon they wit ...*
Negative	Positive	*"An astronaut (Michael Emmet) dies while returning from a mission and his body is recovered by the military. The base where the dead astronaut is taken to becomes the scene of a bizarre invasion plan from outer space. Alien embryos inside the dead astronaut resurrect the corpse and begin a terrifying assault on the military staff in the hopes of conquering the world," according to the DVD sleeve's ...*
Negative	Positive	*Lovely music. Beautiful photography, some of scenes are breathtaking and affecting. But the dramatic tension is lost in a film that is so poorly edited it is hard to know what exactly is going on. At times, the dialogue is incomprehensible. Then there is Richard Gere. He's supposed to be a factory worker who gets into trouble and gets work on a farm. We see dozens of farmhands sweaty and dirty in ...*
Negative	Positive	*This is about some vampires (who can run around out in the sunlight), that are causing some problems down in South America. Casper Van Dien is sent in with his team of commandos to investigate. The movie opens with Van Dien & Co. walking through the jungle, and there's this huge black guy who just absolutely, positively cannot act. He speaks all his lines as if he's reading them off the cue-cards ...*
Positive	Negative	*Sex, drugs, racism and of course you ABC's. What more could you want in a kid's show!¡br /¿¡br /¿————————————————————-————————————————————-¡br /¿¡br /¿"User Comment Guidelines ¡br /¿¡br /¿Please note there is a 1,000 word limit on comments. The recommended length is 200 to 500 words. The minimum length for comments is 10 lines of text. Comments which ar ...*
Negative	Positive	*I used to always love the bill because of its great script and characters, but lately i feel as though it has turned into an emotional type of soap. If you look at promotional pictures/posters of the bill now you will see either two of the officers hugging/kissing or something to do with friendships whereas promotional pictures of the bill a long time ago would have shown something to do with crim ...*
Positive	Negative	*i was having a horrid day but this movie grabbed me, and i couldn't put it down until the end... and i had forgotten about my horrid day. and the ending... by the way... where is the sequel!!!¡br /¿¡br /¿the budget is obviously extremely low... but ... look what they did with it! it reminds me of a play... they are basically working with a tent, a 'escape pod', a few guns, uniforms, camping gear, ...*
Positive	Negative	*Great cult flick for MST-3K types: Richard Boone is a mess – bad hair, arthritis, even his dark glasses aren't right; about as good as a bad dino-flick can get... actually, that charging saber-toothed Styracosaurus was pretty cool – maybe Spielberg should take a couple of notes from that one. ...*
Negative	Positive	*This is, without a doubt, the most hilarious movie I've ever seen. Seriously, if the makers of this movie are ever discovered, they'll put guys like Jim Carrey out of a job. Rent "Jack-O" tonight! Believe me, you won't regret it! ...*

Table 9: Ten misclassified IMDb samples under LoRA Fine-Tuning.

References

[1] Zhou, D. (2025). *DSA4213 Lecture 6: In-context learning and Instruction Finetuning.* AY25/26 Semester 1 course lecture slides, National University of Singapore.