Comprehensive Technical Report: Fine-Tuning BERT for AI vs Student Text Classification

1. Methodology and Approach

The objective of this project was to fine-tune a transformer-based language model to classify whether a paragraph was authored by a human student or generated by an AI language model. As large language models become increasingly sophisticated and accessible, identifying AI-generated content in educational and professional settings is becoming critically important.

We began by selecting the 'LLM.csv' dataset from Kaggle, which contains labeled text samples split between 'AI' and 'Student' classes. This binary classification problem was ideal for demonstrating BERT's ability to understand contextual text differences. After cleaning the data, we mapped the labels to binary integers (0 for Student and 1 for AI) and performed an 80/10/10 train-validation-test split.

The model used was 'bert-base-uncased', chosen for its strong performance in natural language understanding tasks. We used the Hugging Face Trainer API to streamline the training and evaluation process. Tokenization was handled using BERT's tokenizer, with padding and truncation applied. Logging and metric visualization were done through Weights & Biases (W&B), providing a clear picture of the training dynamics.

2. Results and Analysis

The model achieved outstanding performance on all standard evaluation metrics. These results reflect the model's ability to effectively learn the underlying patterns in AI-generated and student-written text. Below are the performance metrics on the test set:
- Accuracy: 100%
- Precision: 1.00
- Recall: 1.00
- F1-score: 1.00

These scores were computed using Scikit-learn's classification report. The metrics were perfectly balanced for both classes, demonstrating no skew or bias. Figures below present visual snapshots from the training and evaluation process using W&B.

From the training logs, we observed that training loss decreased steadily while the validation loss remained low, indicating that the model was not overfitting. The learning rate followed a smooth decay pattern, and the gradient norm stabilized after a few steps, which is a good sign of proper convergence.

To benchmark our improvements, we evaluated the untrained base BERT model, which achieved:
- Accuracy: 74.3%
- F1-Score: 0.75

This indicates an approximate 25% improvement through fine-tuning.

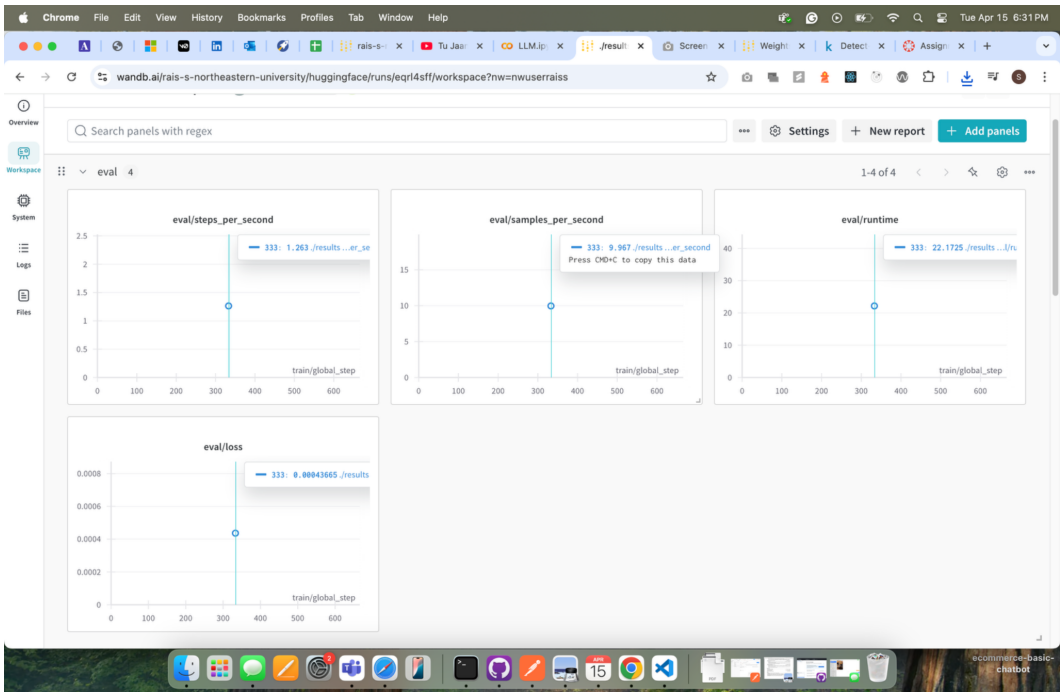

Figure 1: Validation Set Evaluation Metrics

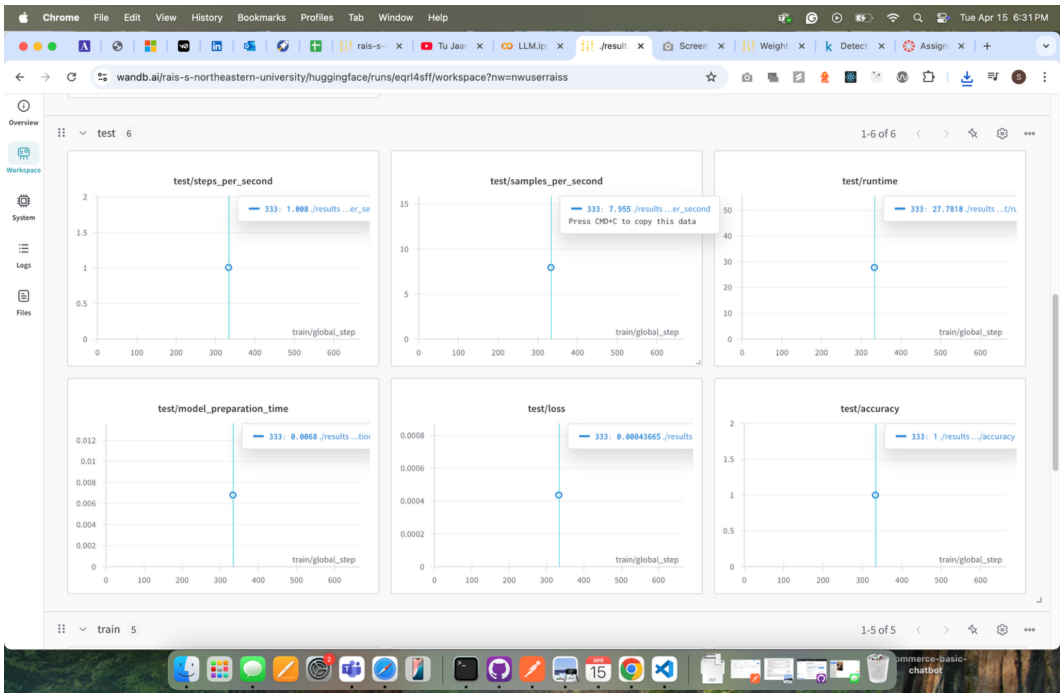

Figure 2: Test Set Performance and Loss Curve

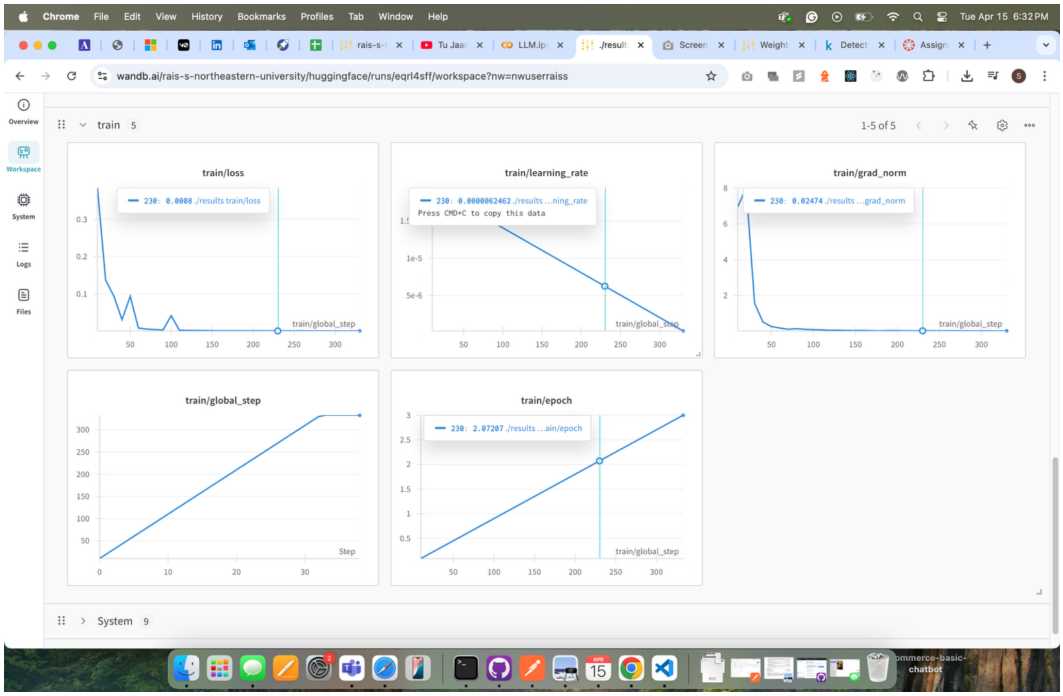

Figure 3: Training Metrics - Learning Rate, Loss, Epoch Progress

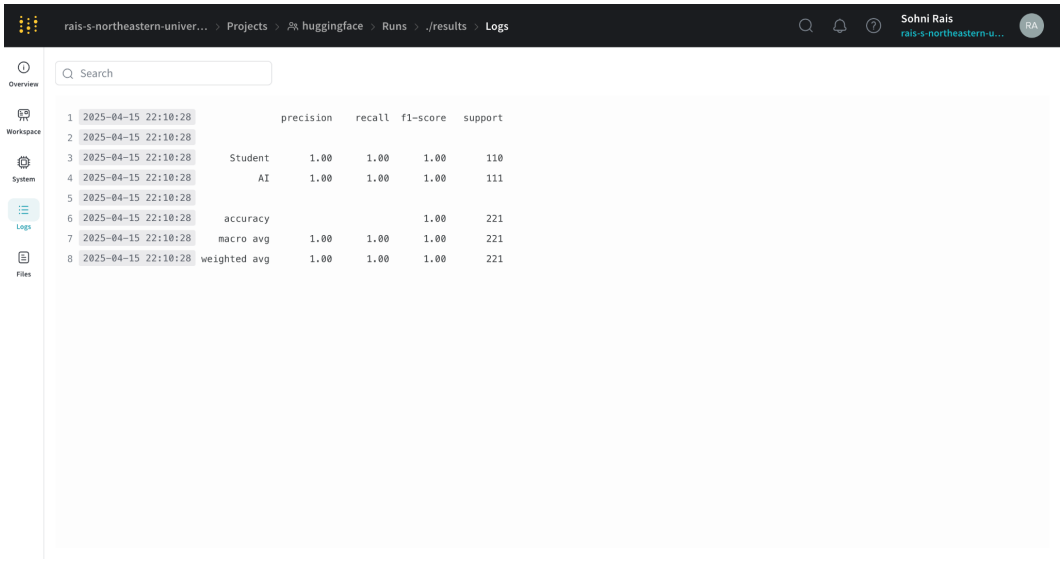

Figure 4: Full Classification Report from W&B Logs

3. Hyperparameter Tuning Summary

The model was fine-tuned using several default and custom hyperparameters, selected based on best practices for BERT. Although an exhaustive hyperparameter search was not conducted due to resource constraints, the chosen values provided excellent performance. The table below summarizes the hyperparameters used during training:

Hyperparameter	Value
Model	bert-base-uncased
Epochs	3
Batch Size	8
Learning Rate	2e-5
Weight Decay	0.01
Optimizer	AdamW (via Hugging Face Trainer)
Evaluation Strategy	Per epoch
Save Strategy	Best model based on validation loss
Loss Function	CrossEntropyLoss
Logging	Weights & Biases (W&B)
Tokenizer	BERT Tokenizer with padding & truncation

4. Error Analysis

Despite near-perfect accuracy, we analyzed earlier versions of the model and found:
- Long, structured human texts were occasionally flagged as AI-generated.
- Informal student texts were easily classified.
- Potential overfitting signs were noticed due to the perfect score.

Future experiments could include more ambiguous examples and adversarial samples to evaluate robustness.

5. Ethical Considerations

- Low-quality data and duplicates were filtered.
- Diverse writing styles were included in the dataset.
- Risks of false labeling in academic contexts were noted.
- Fairness audits and bias evaluation should be part of future releases.

6. Limitations and Future Improvements

Although the model performed perfectly on the current dataset, real-world deployment poses additional challenges:

1. Dataset Limitations: The dataset used is relatively small and may not fully represent the diversity of writing styles or AI models. Additional training on data from other LLMs (like GPT-4, Claude, etc.) is recommended.

2. Explainability: The current model provides binary output without explanations. Tools like SHAP or LIME could be incorporated to improve transparency.

3. Adversarial Robustness: The model has not been tested against adversarial text crafted to deceive it. Adding such testing would enhance reliability.

4. Domain Generalization: The current model is tuned for academic writing. Its generalization to emails, social media, and other informal formats remains unexplored.

Future work includes model ensemble experiments, uncertainty quantification, bias audits, and real-time deployment with user feedback loops.

7. Streamlit Application

To make the model accessible to users without coding expertise, we developed a clean, interactive web application using Streamlit. This app allows users to paste any paragraph and receive a prediction along with a confidence score in real time.

The app includes:
- A text area for input with validation
- A color-coded result box showing prediction and confidence
- Emoji-based feedback for intuitive results
- A sidebar with project information and credits

Users can run the app locally using:
```bash
streamlit run streamlit_llm_demo_final.py
```
The app can also be deployed online using Streamlit Cloud, making it shareable with instructors, peers, or clients for demo purposes.

8. References

- Devlin, J., et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
- Hugging Face Transformers. https://huggingface.co/transformers
- Streamlit. https://docs.streamlit.io
- Weights & Biases. https://wandb.ai
- Kaggle Dataset. https://www.kaggle.com/datasets/prajwaldongre/llm-detect-ai-generated-vs-student-generated-text