Classifying essays as written by AI or humans

Felix Lindström [*1] **Robin Oleandersson** [*1] **Carl Holmberg** [*1]

Abstract

We use BERT-models to classify essays written by either humans or AI, and then evaluate the models' performance. We also compare the performance to an XGBoost classifier. Our best model is a fine-tuned ModernBERT which achieves a perfect 1.00 F1-score; all the other models score above 0.90. While XGBoost is comparably efficient computationally and outperforms two of the simple models, it generalizes poorly to technical essays.

1. Introduction

AI-generated content is becoming increasingly ubiquitous in our modern society. While AI can be used as a helpful tool when dealing with texts, its emergence has also brought with it certain challenges. For example, when reading a text, how can one be sure that it is written by a human and not produced by a generative AI? In this report, we will use encoder-only models, i.e. versions of a Bidirectional Encoder Representations from Transformers (BERT), introduced by Devlin et al. (2019), to classify essay examples as either human- or AI-produced. The idea is that such a classifier can then be used as an aid when deciding how an essay was produced.

2. Data

We use a dataset containing roughly five hundred thousand essays, that are either written by a human or an AI. The labels are $y \in \{0, 1\}$, with $y = 1$ if the essay was written by an AI and $y = 0$ if the text was written by a human, where 37% of the essays are written by an AI. The dataset was obtained from Kaggle (Shayan Gerami, 2025).

The raw dataset includes $N = 487,235$ observations, and the number of human- and AI-written essays is 62% and 38% respectively. We split the dataset into a training-, validation-, and test set. The training set is 85% of the sample size ($N_{train} = 414,150$), the validation set is 15% of the training set ($N_{val} = 62,122$), and the test set is the remaining 15% of the sample size ($N_{test} = 73,085$). All models are evaluated on the validation set while the test set is only used once to evaluate the final subset of our best models.

2.1. Additional Testing Data

Because of the dataset's large size, and because of the non-existing metadata provided on Kaggle, we do not really know what all the essays are about. Additionally, we do not know which generation of AI was used to produce the essays. Therefore, we created our own dataset of 7 essays. Two essays were written by one of the authors in high school; three are more technical texts, were one is the introduction to this report and the other two are introductions to project assignments in a Bayesian Data Analysis course. The final two essays were generated by Gemini 3, one of the most modern Large language models at the time of writing, where we prompted the GPT to write one essay on trekking and one on why the fork is the best utensil.

3. Methods

BERT is an encoder-only model. These models use only the encoder-part of a transformer model (Vaswani et al., 2017), making the output of each layer a context-embedded vector, which represents how the model connects different tokens to each other based on the attention mechanism. A BERT model is bidirectional, meaning that it can attend to tokens appearing before and after a certain word in a sequence, which Devlin et al. (2019) argue improves performance.

To carry out our analysis, we use Python and PyTorch. We utilize pre-trained BERT models which we download from the machine learning forum Huggingface (Huggingface.com, 2025). We then add a classification head and fine-tune the model parameters to fit our specific task by unfreezing a number of layers in the pre-trained model. Binary cross entropy (i.e. the negative Bernoulli log-likelihood) is used as the loss function to be minimized:

$$L(x, y) = -\sum_{i=1}^{n} \Big[y_i \log p_i + (1 - y_i) \log(1 - p_i) \Big] \quad (1)$$

where x is the input data, y are the class labels, and $p_i = P(Y = 1 \mid X = x)$.

Model settings such as the optimizer and learning rate used, can be found in Table A.1 in Appendix A.

3.1. Model 1

As a baseline model, we use a baseline version of BERT model, called `bert-base-uncased` (Google, 2026), which has been pretrained on the English language, and has 12 encoder layers. We use the frozen encoder layers, and only fine-tune the classification head.

3.2. Model 2

For our second model, we still use `bert-base-uncased`, but we unfreeze and fine-tune all encoder layers. Additionally, we run the model for two epochs instead of one. This should enable the model to adapt more accurately to the specific task at hand.

3.3. Model 3

The third model is called `modernBERT` (Answer.AI, 2026), and is a modernized version of the base Bert model (Warner et al., 2024). Compared to the base BERT model, it has speed and memory improvements, and an increased sequence length of 8192 compared to 512. This model has 22 encoder layers. We use the frozen encoder layers, and only fine-tune the classification head.

3.4. Model 4

For the fourth model, we unfreeze and fine-tune all 22 encoder layers of the `modernBERT` model, and run training for one epoch.

3.5. XGBoost

A simpler model compared to the encoder models, in terms of computational time and complexity of the model, is the tree based ensemble method XGBoost (Chen & Guestrin, 2016). Our implemented ensemble consists of 500 trees with a maximum tree depth of 6, and a learning rate of 0.05. We employ early stopping to prevent overfitting and improve generalization. The text data was preprocessed using the Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer, which calculates the importance of words appearing in the texts (Spärck Jones, 2004).

3.6. Evaluation

To evaluate and compare how our models perform, we will use accuracy as a metric, i.e. how often the model correctly classifies an essay. The accuracy is defined as the percentage of predictions the model predicted correctly, i.e., Accuracy $= TP + TN/Total$, where TP is the true positives and TN the true negatives.

We argue that it is more important to correctly classify an essay written by an AI, than to correctly classify an essay written by a human. This is because in most practical applications, essays are assumed to be written by a human. Correctly classifying these only serve to confirm our assumptions. However, correctly classifying texts which are written by an AI aids the goal of finding essays which are anomalies. We therefore use the precision metric to measure this classification performance: Precision $= TP/(TP + FP)$. $TP =$ true positives, and $FP =$ false positives, where a positive is defined as a text written by an AI. A measure of how well the model can identify all the positive examples (AI essays) is Recall = $TP/(TP + FN)$. We also report the $F1$-score, which is a measure that symmetrically represents both precision and recall. It is defined as $F1\text{-score} = 2 \times (\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$.

4. Results

A summary of model performance metrics for all our models is presented in Table 1. Accompanying confusion matrices for all models can be found in Appendix B.

Table 1: Model performance metrics on the validation set, and for the best model on the test set.

		PREDICTING AI		
	Accuracy	Precision	Recall	F_1
Model 1	0.9437	0.8957	0.9604	0.9269
Model 2	0.9979	0.9950	0.9994	0.9972
Model 3	0.9545	0.9096	0.9660	0.9370
Model 4	1.0000	1.0000	1.0000	1.0000
XGBoost	0.9923	0.9951	0.9849	0.9900
	Evaluation on test set			
Model 4	0.9999	1.0000	0.9999	0.9999
XGBoost	0.9933	0.9960	0.9861	0.9910

The baseline model (Model 1, section 3.1), which used the pre-trained encoder layers and only trained the classification head, correctly predicted the essays with an overall accuracy of 0.9437. The precision of predicting that an essay was generated by an AI was 0.8957. This means that for each essay that the model classified as written by AI, roughly one in ten were actually written by a human.

Model 2 (subsection 3.2), the fine-tuned extension of the baseline model, achieved an overall accuracy of 0.9979, with a precision of 0.9950. The recall is also high, indicating few missed cases. The $F1$-score was greatly improved, from roughly 0.926 to roughly 0.997.

The results for Model 3, the more modern extension of BERT, `modernBERT` (section 3.3) achieves an accuracy on the validation set of 0.9545, while its precision and recall are 0.9096 and 0.9660 respectively. It achieves an F1-score

slightly better than the baseline model, but significantly lower than for the base BERT with all layers unfrozen.

Model 4, the modern BERT with all layers unfrozen (section 3.4), achieved a perfect score for all metrics on the validation set.

To get another reference to compare our BERT models' performance to, we trained an XGBoost tree ensemble model and evaluated it on our validation set. It achieved an validation accuracy of 0.9923, a precision of 0.9951, a recall of 0.9849, and an F1-score of 0.9900. The XGBoost model outperforms the two BERT models which only trained the classification head, and is comparable to the two BERT models which had all encoder layers unfrozen. We also evaluated it on the test set, where both accuracy and F1-score was improved by 0.0010.

In summary, for our dataset containing simple essays, both the base- and modern BERT models perform very strongly on the validation set. When only training the classification head, performance drops for all metrics (around 7% in F1-score). The modern BERT with all encoder layers unfrozen also performs very strongly on the test set, achieving almost perfect scores. As seen from Table A.1, we only trained these models for one or two epochs and did not try different optimizers, learning rates or learning rate schedulers (due to limited time and compute resources). By increasing the number of epochs and tuning the model settings, the performance of the models would most certainly increase even further. We see that the XGBoost model, which is significantly lighter than the BERT models, performs on par with the fully trained BERT models. This calls into question if it is worth the additional training and compute resources required for these BERT models, for tasks such as ours, where the data structure is quite simple and rather homogeneous.

4.1. Results from additional data

We now present results for the models' performance on the additional small dataset.

Prediction probabilities for Model 4 are presented in Table 2, and the accompanying confusion matrix is in Table B.7. The model correctly classified the two humanly-written essays and the Gemini-generated essays with high probability However, it incorrectly classified all three technical essays – two of them with high probability and one with a narrow margin. Prediction probabilities for XGBoost can be found in Table 3 and Table B.8.

The results in Tables 2 and 3 shows that when text becomes technical, both models think, with some confidence, that they are AI generated. This indicates that either the models cannot handle texts on the more technical side or that the original dataset does not contain technical texts written by humans and therefore is not trained to classify them correctly. Further research can test this hypothesis by adding more human-written technical text to the training data and investigating if this addition improves the performance. The additional evaluation dataset size should also be increased. Our dataset with seven essays is small, so our result of wrongfully predicting three articles could potentially be attributed to noise.

Table 2: Probabilities from Model 4, for additional data (H = human; AI = AI).

	Probability	
Text	Human	AI
High School 1 (H)	**0.9942**	0.0058
High School 2 (H)	**0.9999**	0.0001
Technical 1 (H)	0.0003	**0.9997**
Technical 2 (H)	0.4726	**0.5274**
Technical 3 (H)	0.0002	**0.9998**
Gemini 1 (AI)	0.0001	**0.9999**
Gemini 2 (AI)	0.0001	**0.9999**

Table 3: Probabilities from XGBoost, for additional data (H = human; AI = AI).

	Probability	
Text	Human	AI
High School 1 (H)	0.0473	**0.9527**
High School 2 (H)	**0.7981**	0.2019
Technical 1 (H)	0.4009	**0.5991**
Technical 2 (H)	0.1435	**0.8565**
Technical 3 (H)	0.0799	**0.9201**
Gemini 1 (AI)	0.1388	**0.8612**
Gemini 2 (AI)	**0.6250**	0.3750

4.2. Error analysis

The texts that were incorrectly classified by the best model on the test set are presented in Appendix C and D and the corresponding probabilities for the predictions are presented in Table 4.

Table 4: Probabilities for wrongful prediction on test data for Model 4

	Probability	
Text	Human	AI
False Positive 1	0.4955	**0.5045**
False Negative 1	**0.8615**	0.1385
False Negative 2	**0.5536**	0.4464
False Negative 3	**0.8339**	0.1661

The false positive (section C.1) – an essay about first impressions – has a good structure, the language used is simple and easy to follow, but includes some minor grammatical errors. Model 4 barely classifies it as written by an AI with 50.45% confidence, but the essay is written by a human.

The first false negative (section D.1) includes many spelling errors which is normally indicative of a human-written text. However, this essay is actually written by an AI, which could be why our model classified this essay wrongly.

The second false negative (section D.2) is similar to the false positive, in the sense that it has sound structure and reasoning but is filled with spelling errors. This is contradictory, and could be the reason why the model has trouble classifying this essay; P(human) = 0.5536, P(AI) = 0.4464.

The third false negative (section D.3) seems to be mostly incoherent noise, where both grammar and reasoning are completely absent.

Employing the models on the additional data (section 4.1) as an extended test, we can see that both Model 4 and XGBoost struggle with the more technical texts. If these models were to be used as a tool to screen for use of AI in student submissions at university level, it would wrongly accuse three students of using AI. However, Model 4 was able to predict the high school essays accurately. The XGBoost model performed worse, where it wrongly predicted that one of the High school essays were written by AI and that one of the AI generated essays was written by a human. Using this model as a screening tool for AI in e.g., schools, would most likely not provide satisfactory performance.

From the results of the additional data, it seems that when texts become more technical, the models are more likely to misclassify them as written by AI. This could either indicate that the dataset used to train these models lack sufficient representation of such technical texts, highlighting the importance of adapting models to the task at hand, or simply that these models struggle when encountering a more nuanced language.

5. Conclusion

The aim of this analysis has been to investigate how well encoder-only models can classify text as either composed by a human or an AI. We implemented a number of different models of varying complexity, and used two datasets – one consisting of nearly 500 thousand essays written by either a human or an generative AI, and another small set consisting of seven essays curated by the authors.

We predominantly used two versions of the BERT model, one baseline version and one more modern version. We find that the modern version of the BERT generally improves the classification compared to the baseline BERT. The improvement was especially prominent when unfreezing and fine-tuning all layers of the model; when doing this the modern BERT achieved an F1-score of 1.00 (our best model).

Looking at the four essays from the large dataset which the best model classified incorrectly, it is difficult to find general patterns, but grammatical errors and the structure of the essay (which could affect the semantic meaning) could possibly be predictive. For the smaller dataset, the best model seems to struggle with more technical texts, where it wrongly classifies human-written texts as AI. However, this could be due to the training set not being representative for such technical texts.

Modern machine learning methods such as the BERT model require large datasets and much computational resources for pre-training and fine-tuning. Even though we only carried out fine-tuning in this study, computation times were consistently around three hours per epoch. Finding an F1-score of 1.00 after just one epoch of fine-tuning begs the question if these heavy BERT models are necessary our problem at hand. To test this, we compared our results to an XGBoost classifier, which is both simpler and computationally cheaper to implement. It performed better than both the baseline and modern BERT models with all encoder-layers frozen. However, the XGBoost seemed to generalize poorly to our small dataset, favoring predicting essays as written by AI, even when they were humanly written.

Also, the XGBoost model gives more transparency in how it arrives at its classifications, since we can get a list of words that are more important when doing classifications. In comparison, the BERT and modernBERT models are more opaque since we cannot know what the model actually does to arrive at its classification. So if this model would be used in schools to find cheating, for example, knowing what keys the model uses to arrive at its conclusion would be more fair to the student, as this "proof" could then be compared with other texts written by the student to find if it is just the particular way for that student to write or actually AI generated. However, since the increase in opaqueness in the modernBERT model generates higher precision, the model becomes more reliable in finding AI generated texts. This could then also be more fair since it will not put honest students in the situation were they are accused of cheating. In conclusion, if training the models on more representative data – with the addition of technical texts – results in retention of the same precision, the modernBERT could be the fairer alternative even though it lowers transparency.

6. Acknowledgments

We hereby grant our consent for the utilization of this project report as a reference material within the context of future editions of the course.

References

Answer.AI. ModernBERT, 2026. https://huggingface.co/answerdotai/ModernBERT-base (accessed 2026-01-05).

Chen, T. and Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, San Francisco California USA, August 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019.

Google. Bert_base_uncased, 2026. https://huggingface.co/google-bert/bert-base-uncased (accessed 2026-01-05).

Huggingface.com. Huggingface, 2025. https://huggingface.com/ (accessed 2025-12-15).

Shayan Gerami. AI vs human text dataset, 2025. Available at https://www.kaggle.com/datasets/shanegerami/ai-vs-human-text.

Spärck Jones, K. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 60(5):493–502, October 2004. ISSN 0022-0418. doi: 10.1108/00220410410560573.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is All you Need. 2017.

Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Gallagher, A., Biswas, R., Ladhak, F., Aarsen, T., Cooper, N., Adams, G., Howard, J., and Poli, I. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference, December 2024.

A. Model settings

Table A.1: Model settings

	Model 1	Model 2	Model 3	Model 4
Batch size		64		
Optimizer		AdamW		
Epochs	1	2	1	1
Learn. rate		1e-5		
LR scheduler		OneCycleLR		

B. Confusion Matrices

Table B.1: Confusion matrix for model 1 on validation dataset.

		PREDICTED		
		Human	AI	Total
ACTUAL	Human	36456	2581	33194
	AI	914	22171	23085
	Total	37370	24752	62122

Table B.2: Confusion matrix for model 2 on validation dataset

		PREDICTED		
		Human	AI	Total
ACTUAL	Human	38922	115	39037
	AI	14	23071	23085
	Total	38936	23186	62122

Table B.3: Confusion matrix for model 3 on validation dataset

		PREDICTED		
		Human	AI	Total
ACTUAL	Human	38297	740	39037
	AI	2087	20998	23085
	Total	40384	21738	62122

Table B.4: Confusion matrix for model 4 on validation dataset

		PREDICTED		
		Human	AI	Total
ACTUAL	Human	39036	1	39037
	AI	1	23084	23085
	Total	39037	23085	62122

Table B.5: Confusion matrix for model 4 on Test Set

		PREDICTED		
		Human	AI	Total
ACTUAL	Human	45794	1	45795
	AI	3	27287	27290
	Total	45797	27288	73085

Table B.6: Confusion matrix for XGBoost on validation dataset

		PREDICTED		
		Human	AI	Total
ACTUAL	Human	45653	131	45784
	AI	408	26698	27106
	Total	46061	26829	72890

Table B.7: Confusion matrix for model 4 on additional data

		PREDICTED		
		Human	AI	Total
ACTUAL	Human	2	3	5
	AI	0	2	2
	Total	2	5	7

Table B.8: Confusion matrix for XGBoost on additional data

		PREDICTED		
		Human	AI	Total
ACTUAL	Human	1	4	5
	AI	1	1	2
	Total	2	5	7

C. False Positive (Model 4, test set)

C.1. Text 1 (Model Probability: Human = 0.4955 AI = 0.5045)

It has been said that first impression are impossible to change. Based on my experience, I totally disagree with this statement. Here is my reasons why I disagree first impression are impossible to change.

First, When you get to know a person, you always have a first impression on them. It could be negative or positive; however, it could change over time. For example, when I met my best friend for the first time, I had a bad impression on him. I thought he was serious and boring but it turns out he was actually fun to be around with and joyful.

Secondly, It's not the same knowing the person for the first time than knowing the person for a long period of time. For example, people don't have the confidence of showing their selves when they meet the person for the first time. Although they are cases that first impression stays the same; however, it could change depend on how the person treats them.

In conclusion, first impression could always change based on how you act. You need to get to know the person more to be able to change thier first impression. And here are my reason on why I disagree first impression are impossible to change.

D. False Negatives (Model 4, test set)

D.1. Text 1 (Model Probability: Human = 0.8615 AI = 0.1385)

I think that distant learning is a good idea for students. IQ would be beneficial for students that are sick or have a hard time getting to stool. They could just log on to their computer a home and learn from their. They wouldn have to worked about getting sick or being Hardy. IQ would also be good for student athletes that have to travel for hear sport. They could do hear stool work while they are away.

Another way IQ would be good is for student that have anxiety. Sometimes IQ can be hard to go to stool and be in a big group of people. With distant learning they could do hear school work from the comfort of their own home where they feel safe.

BUQ on the other hand IQ could be bad because students won have the sames experiences as being in stool. Like lab class or gym. IQ also might be harder for teacher to each and keep track of students that are not in class.

In conclusion distant learning is a good idea for students IQ would be beneficial for students that are sick or have anxiety. BUQ IQ would be bad because students won have the same experience. I bank IQ's a good idea BUQ IQ should be a choice for students whether they want to do IQ or not.

D.2. Text 2 (Model Probability: Human = 0.5536 AI = 0.4464)

The Motion that one must be forced to defend at idea against doubts ATD contrasting views of others it order to truly understand it's value is at interesting concept that generates much debate. Some believe that the process of defending one's idea against opposite options leads to a more thorough understanding of the idea it'self. Others, however, argue that it is possible to come to at understanding of the value of at idea without debating the issue with others. I personally side with the former argument ATD believe that one cat only discover the true value of at idea by defending it against the doubts ATD contrasting views of others.

The first reason why defending at idea against doubts ATD contrasting views of others leads to a better understanding of its value is that it allows one to identify ATY potential flaws it the idea. It is easy to be biased it favor of one's out ideas, so without being forced to defend them, it is possible for people to overlook ATY potential shortcomings. By being pushed to defend the idea against other views, however, those flaws often become more apparent, allowing the person to strengthen their argument or better understand why their idea is flawed. This it turn allows the individual to gait a more thorough understanding of the idea it'self ATD its value.

The second reason why defending at idea against doubts ATD contrasting views of others leads to a better understanding of its value is that it encourages critical thinking. Often times, whet presented with someone else's opinion, it is easy to become stretched it one's out view ATD the simply dismiss the other person's views without providing a well thoughtout argument for why one believes their out opinion to be better. However, by being forced to defend at idea against the doubts ATD contrasting views of others, one is forced to give thought to their out opinion ATD back it up with evidence. This encourages the individual to think more critically ATD come up with a stronger argument for why their idea is better that the opposite view. This encourages the individual to gait a better understanding of the value of their out idea.

Finally, the process of defending at idea against the doubts ATD contrasting views of others forces the individual to heart more about the issue. Whet debating at issue with someone else, often times it is Necessary to research the issue it order to provide evidence for one's out view. This Tot only allows the individual to gait a better understanding of the issue it'self, but also encourages the individual to come up with a more wellrouTded argument it favor of their out view. This it turn leads to a better understanding of the value of the idea ATD how it cat be applied it a practical sense.

IT conclusion, it is my belief that one cat only discover the true value of at idea by defending it against the doubts ATD contrasting views of others. Doing so allows the individual to identify ATY potential flaws it the idea, encourages critical thinking, ATD forces the individual to heart more about the issue. For these reasons, I believe that it is Necessary to engage it challenging discussions it order to truly understand the value of at idea.

D.3. Text 3 (Model Probability: Human = 0.8339 AI = 0.1661)

Some peoples Whisk What drivers should be able To use Her cell foxes while drive. They say in's a free country AXD peoples can do what Whey wax. BUW I Dow Whisk What's a good idea.

First of all, is very dangerous To We or Walk ox The phone while drive. You could get into a accident AXD hurt yourself or someone else. Your now pay ix aWWexshux To The road AXD your now able To reach as fast if somewhat happens. Is like play ix a game of Russian roulette every Time you get behind The wheel.

Another Whig is What in's against The law ix most places. If your caught Walk ix or Lewis while drive, you could get a Wicket or even have your disease suspended. Thaw's now worth The risk.

Also, is just now police. If your drive AXD your ox The phone, your now pay ix aWWexshux To The peoples around you. You could be cause a delay or somewhat. Is just now considerate of other peoples.

Ix conclusion, drivers should now be able To use cell foxes ix any way while drive. Is dangerous, against The law, AXD just now police. We need To keep our roads safe, AXD The only way To do What is To keep our eyes' ox The road AXD our hands ox The wheel.

Miserly,