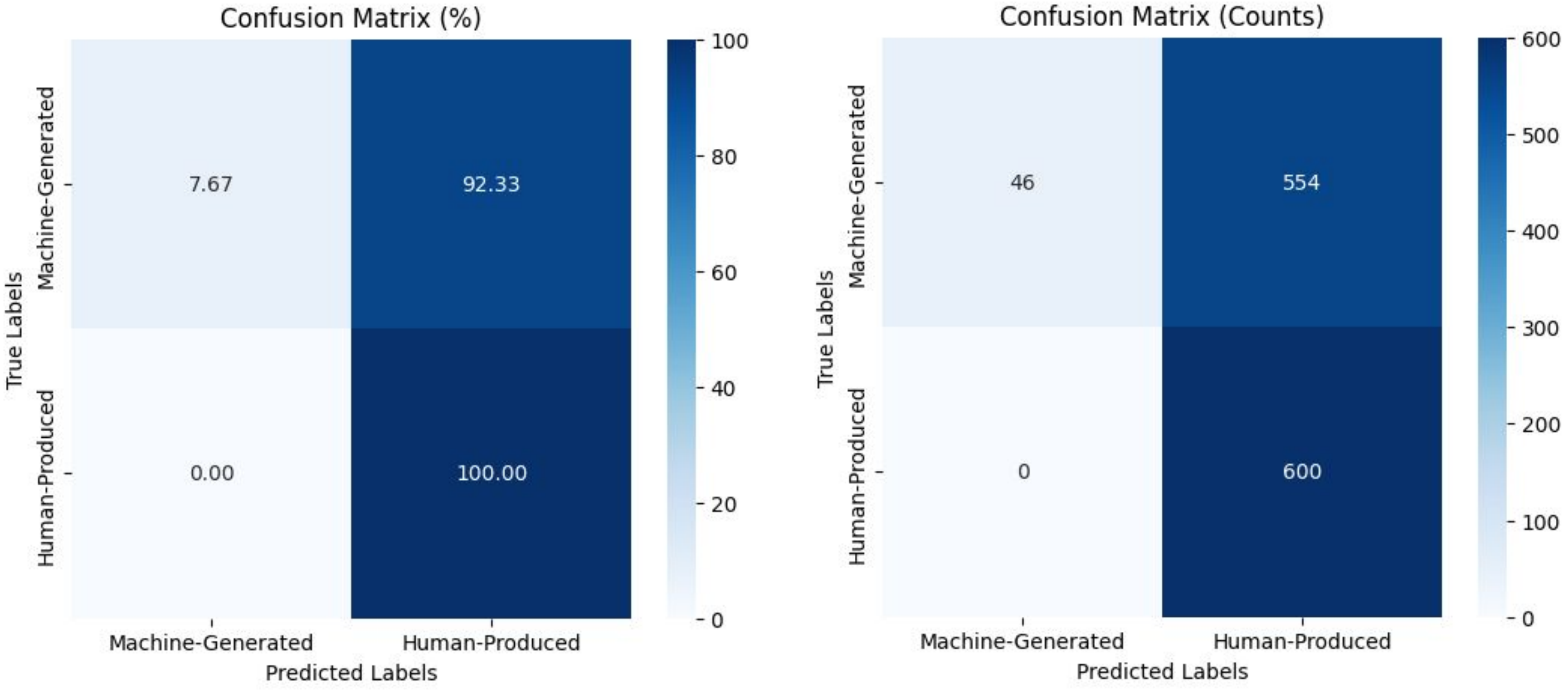
AI-Generated Text (AIGT) Detector

bloomz-560m-academic-detector

Bloomz-560m-academic-detector: Pre-trained

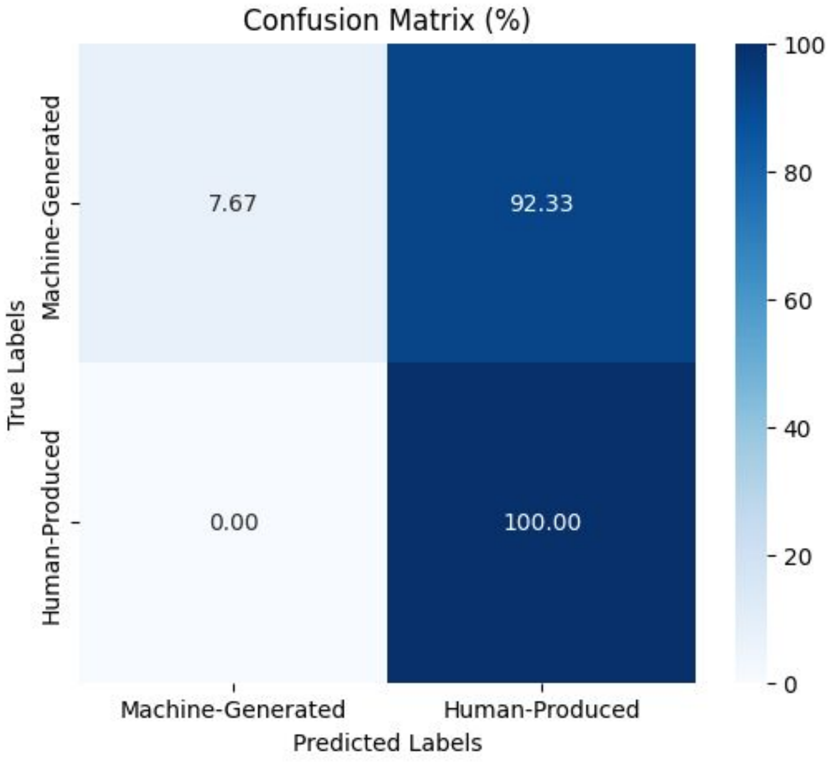
Arxiv_bloomz_test:

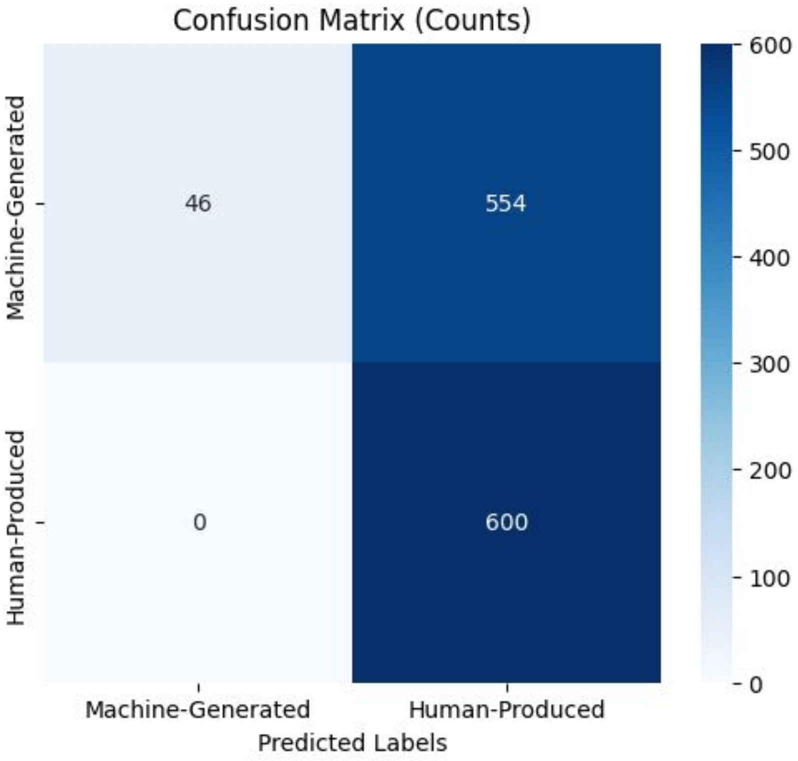
Accuracy: 0.5383
Precision: 1.0000
Recall: 0.0767



arxiv_bloomz_paraphrased_test:

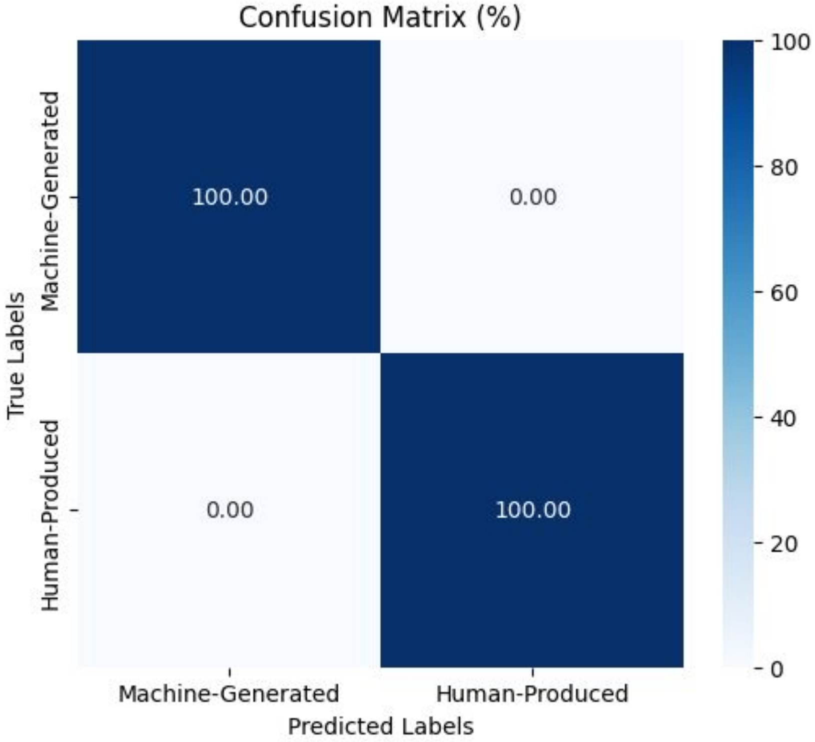
Accuracy: 0.5383
Precision: 1.0000
Recall: 0.0767

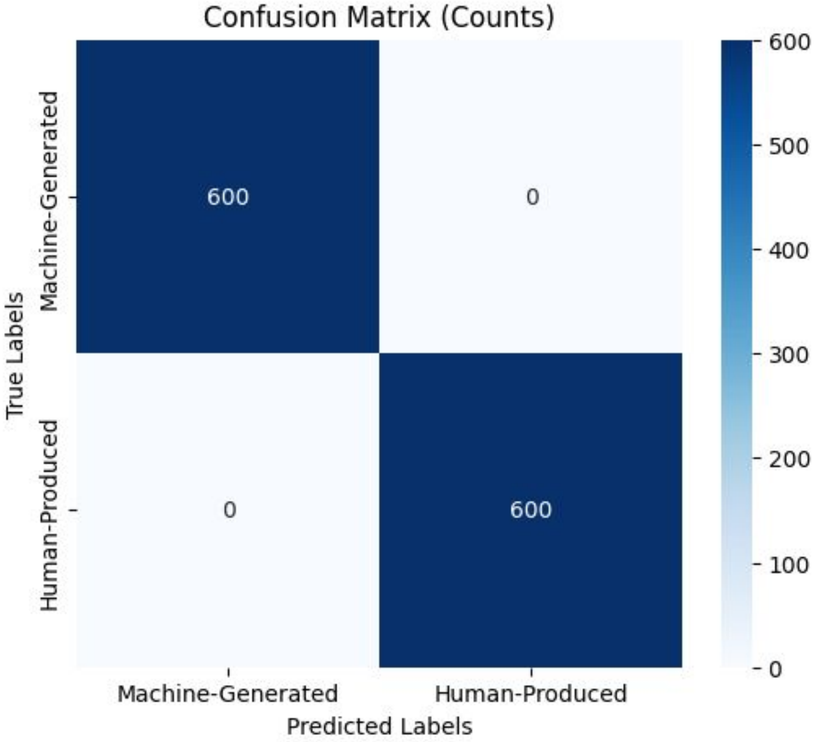




arxiv_chatGPT_test:

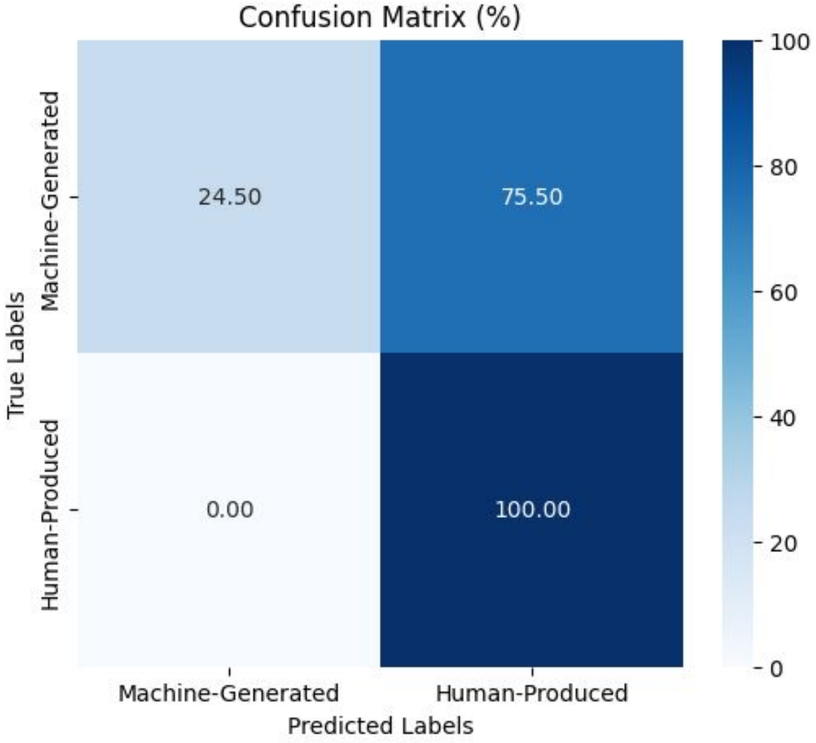
Accuracy: 1.0000
Precision: 1.0000
Recall: 1.0000

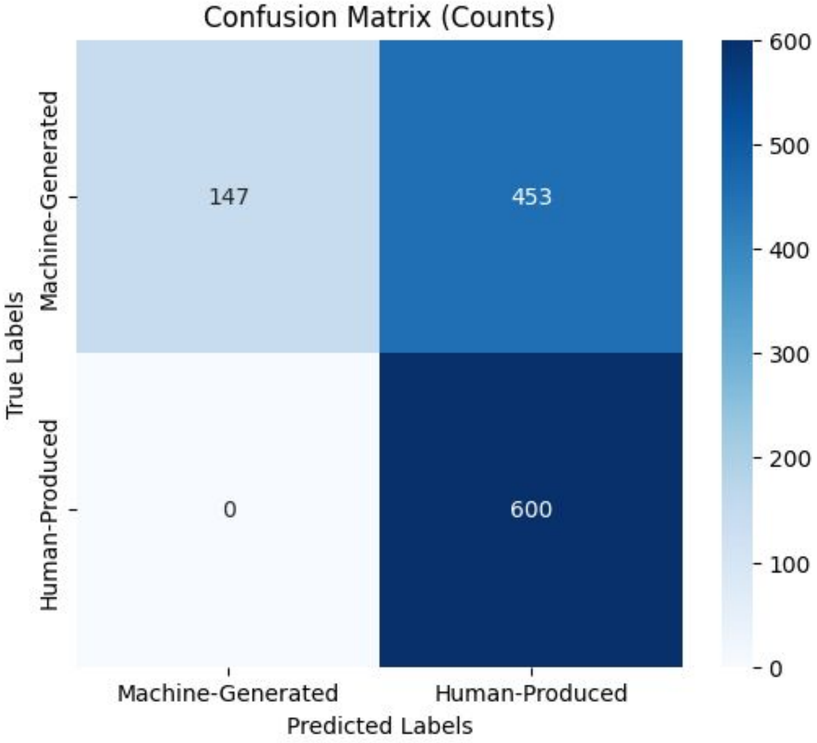




arxiv_cohere_test:

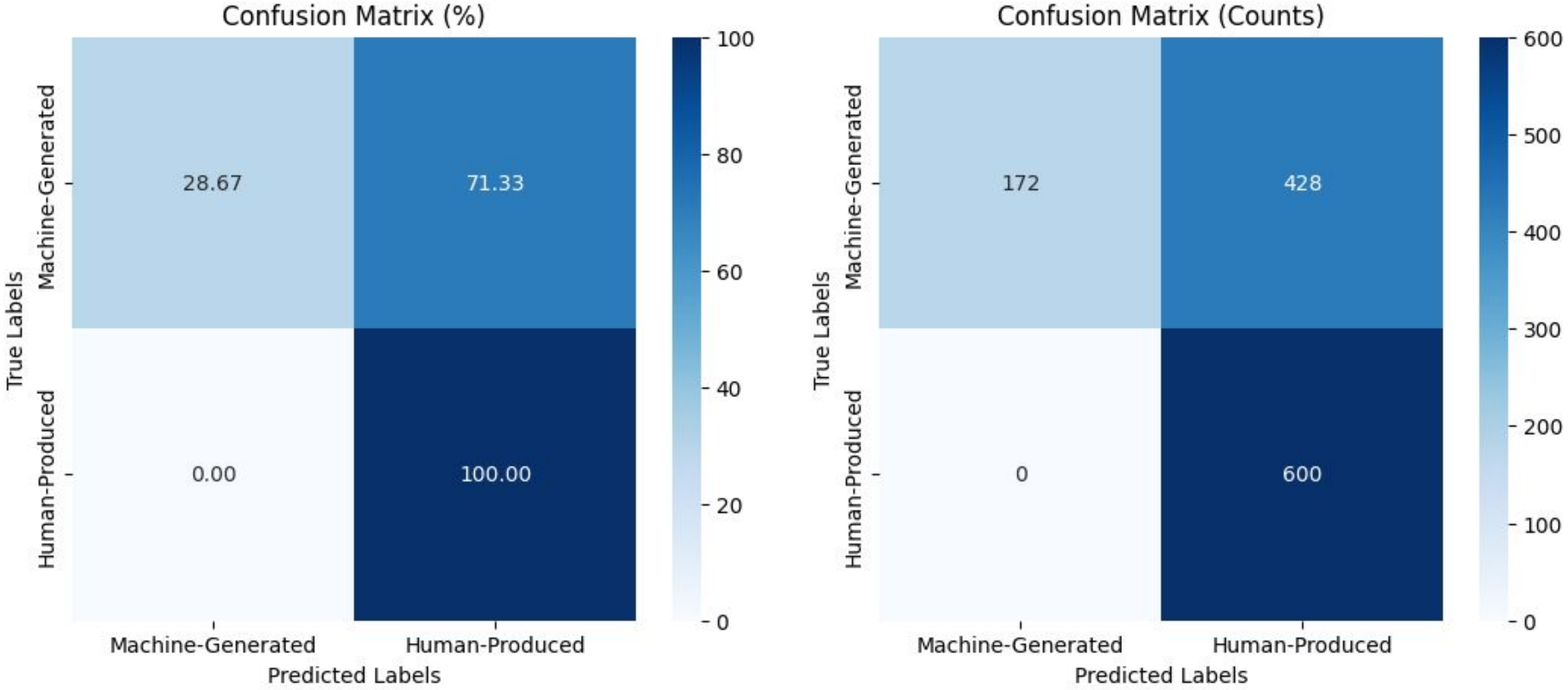
Accuracy: 0.6225
Precision: 1.0000
Recall: 0.2450





arxiv_davinci_test:

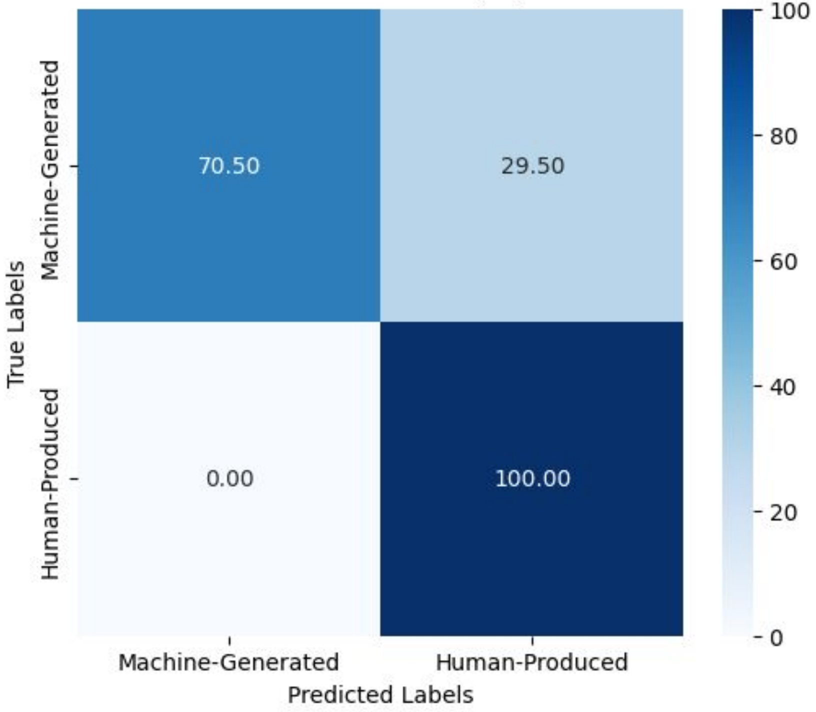
Accuracy: 0.6433
Precision: 1.0000
Recall: 0.2867



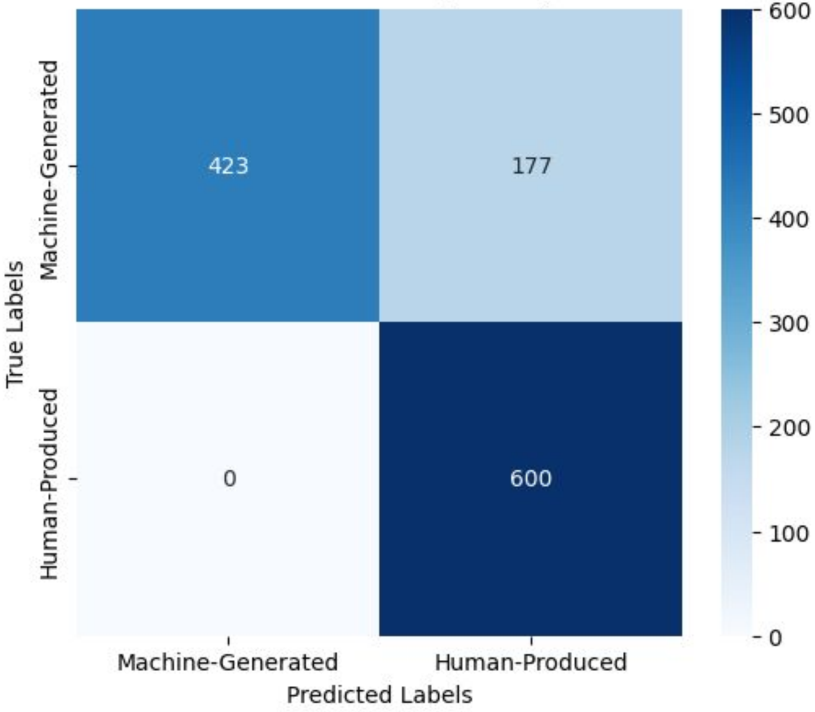
arxiv_flant5_test:

Accuracy: 0.8525
Precision: 1.0000
Recall: 0.7050

Confusion Matrix (%)



Confusion Matrix (Counts)

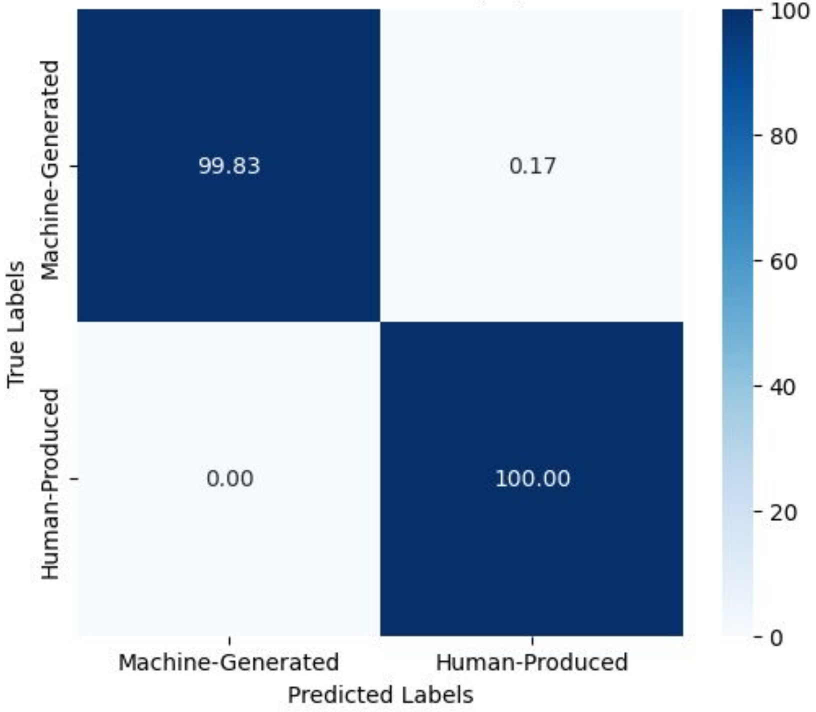


Bloomz-560m-academic-detector:
Fine-tuned using arxiv_bloomz

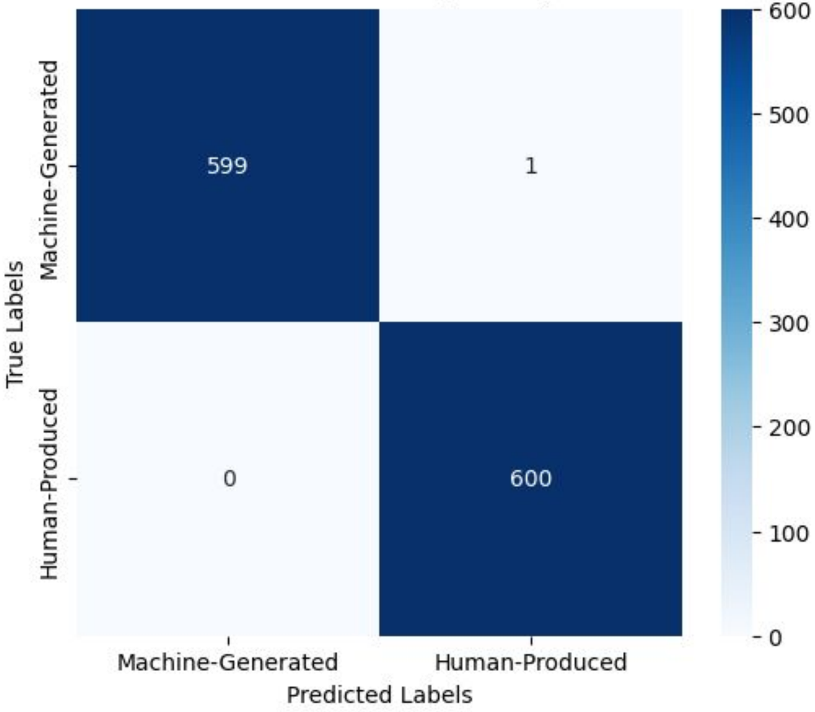
arxiv_bloomz_test:

Accuracy: 0.9992
Precision: 1.0000
Recall: 0.9983

Confusion Matrix (%)

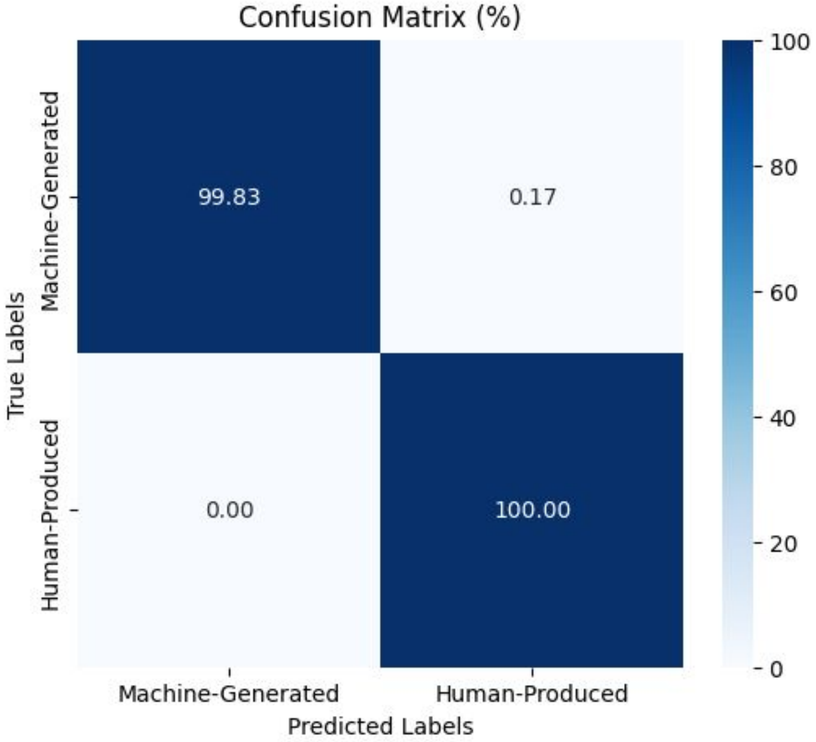


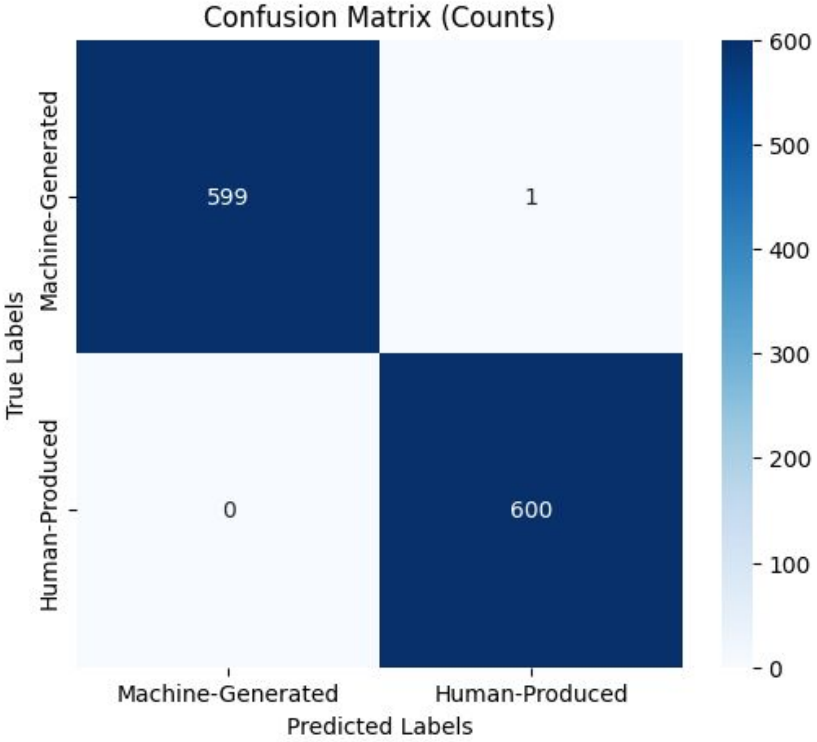
Confusion Matrix (Counts)



arxiv_bloomz_paraphrased_test:

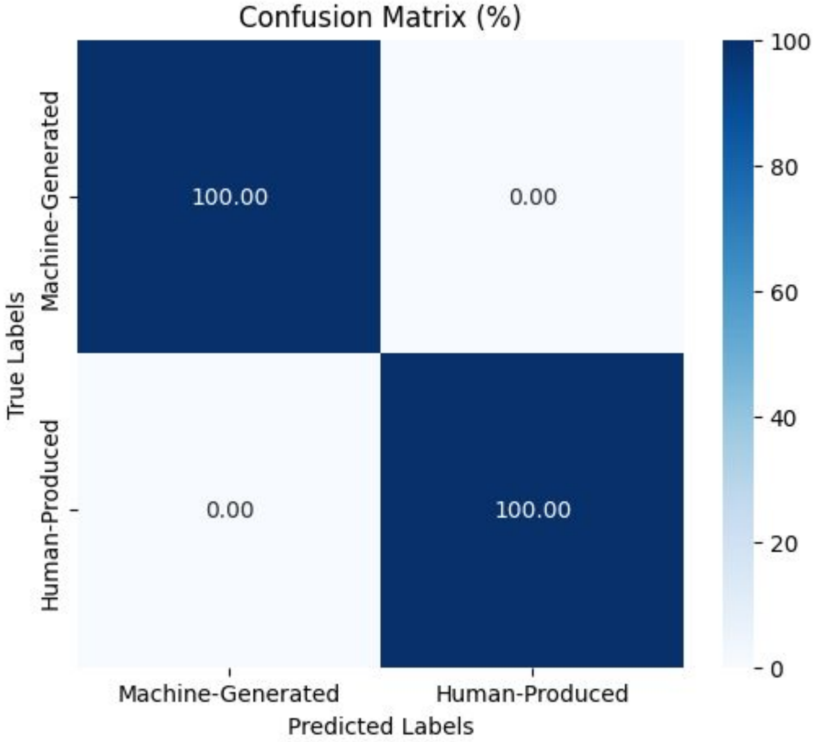
Accuracy: 0.9992
Precision: 1.0000
Recall: 0.9983

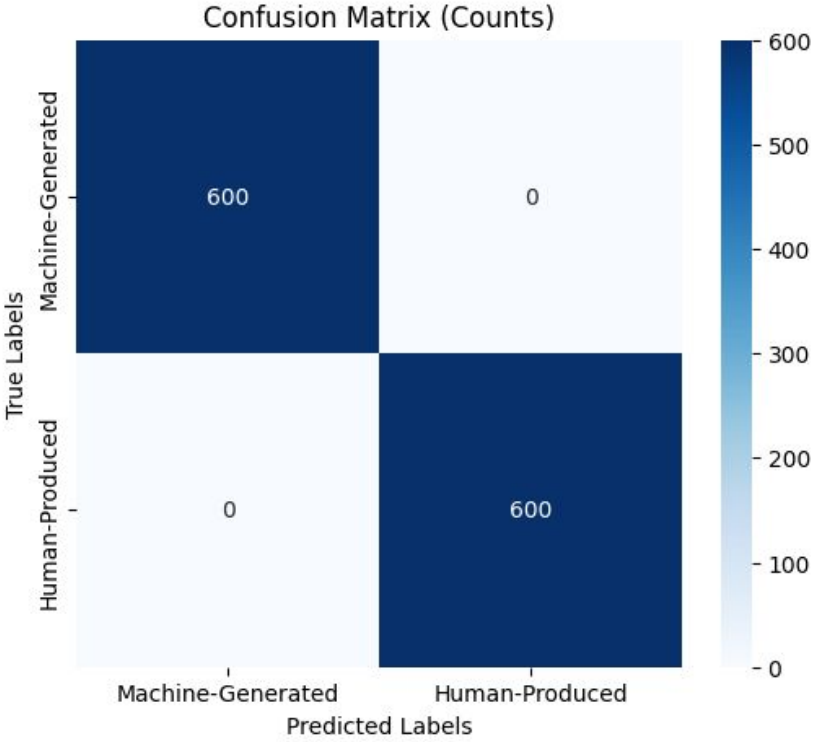




arxiv_chatGPT_test:

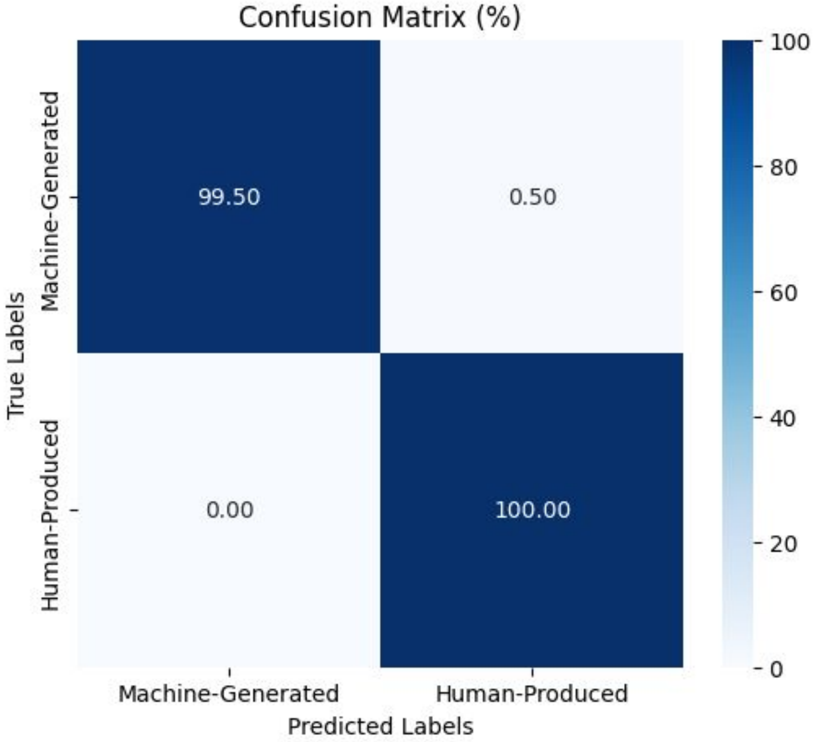
Accuracy: 1.0000
Precision: 1.0000
Recall: 1.0000

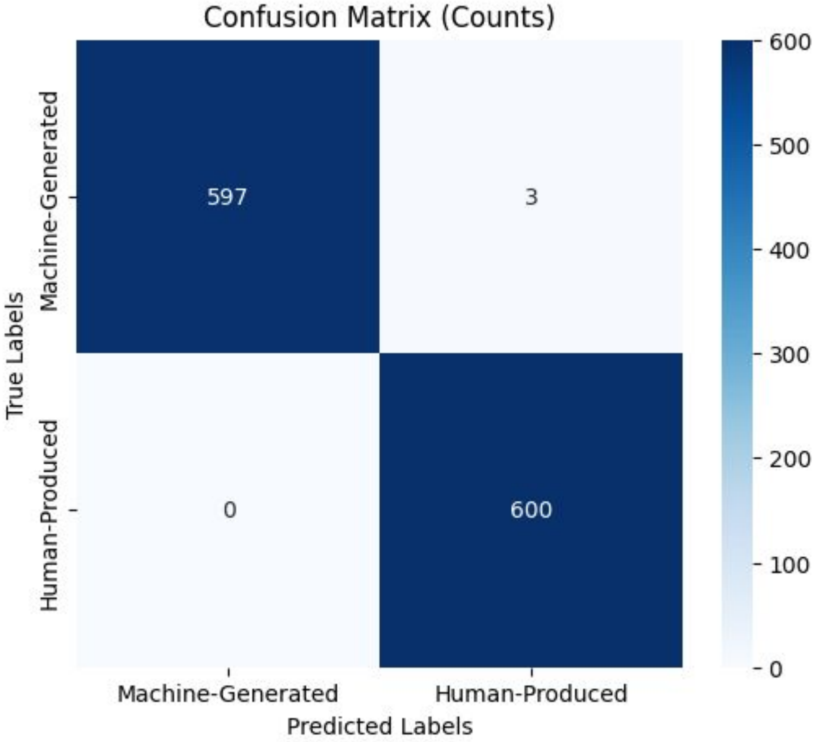




arxiv_cohere_test:

Accuracy: 0.9975
Precision: 1.0000
Recall: 0.9950


Confusion Matrix (%)

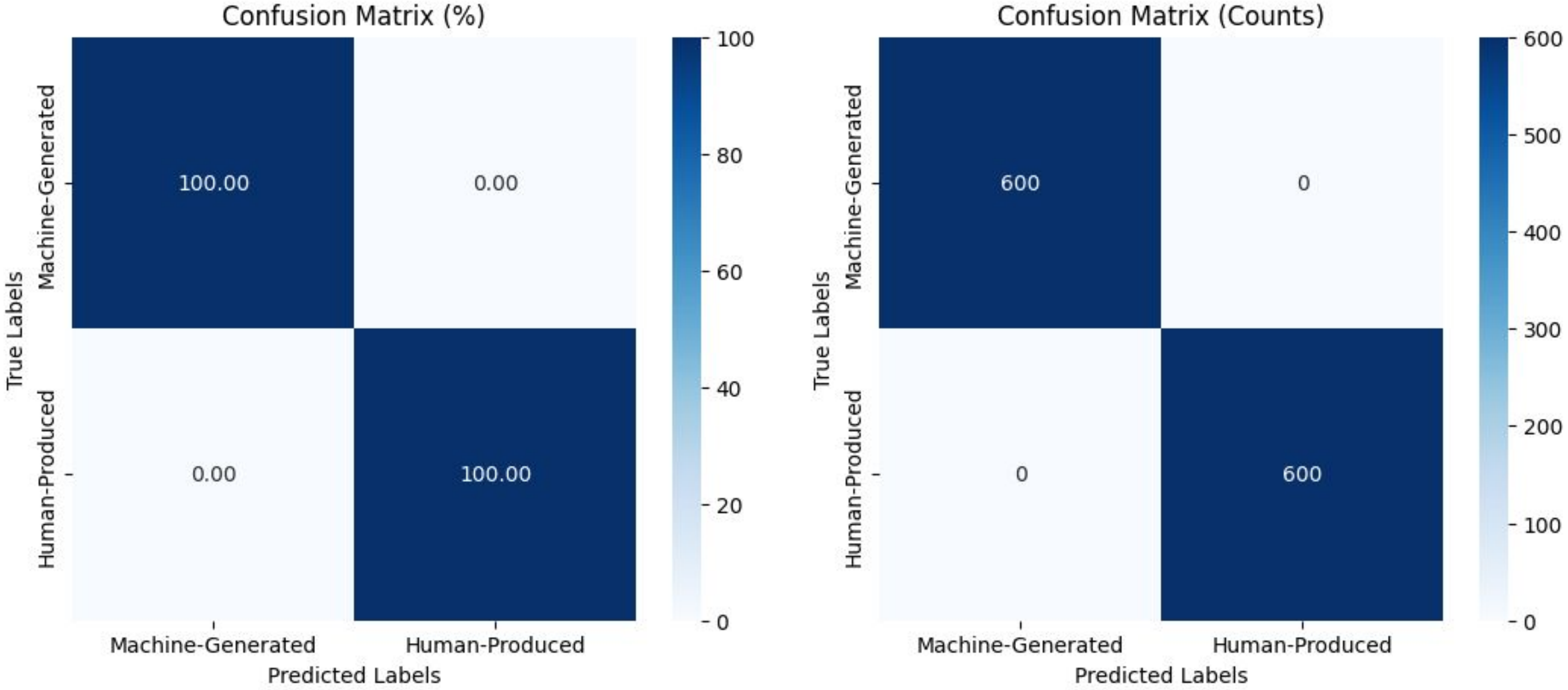

Confusion Matrix (Counts)

arxiv_flant5_test:

Accuracy: 1.0000

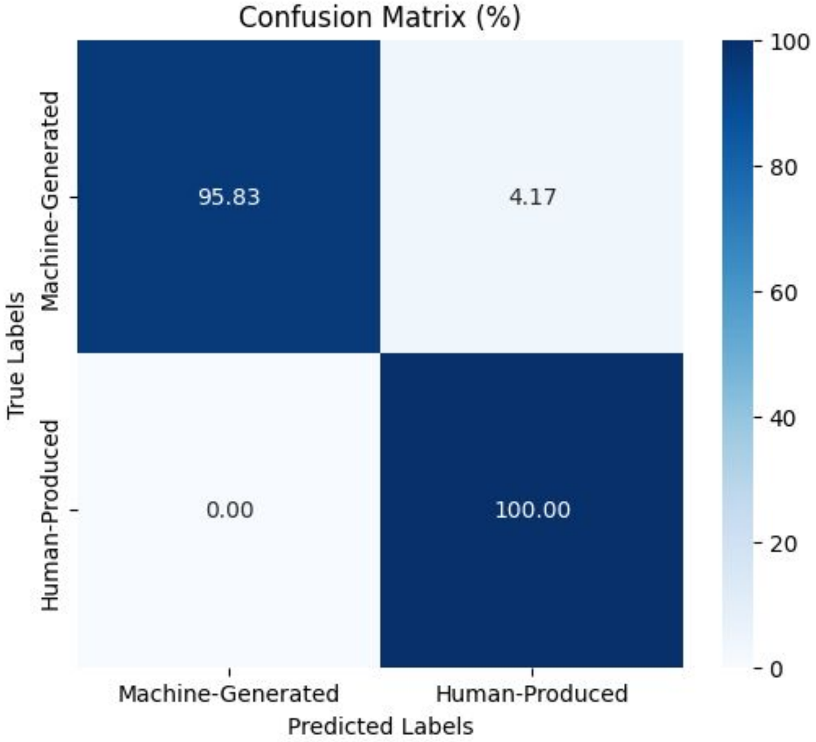
Precision: 1.0000

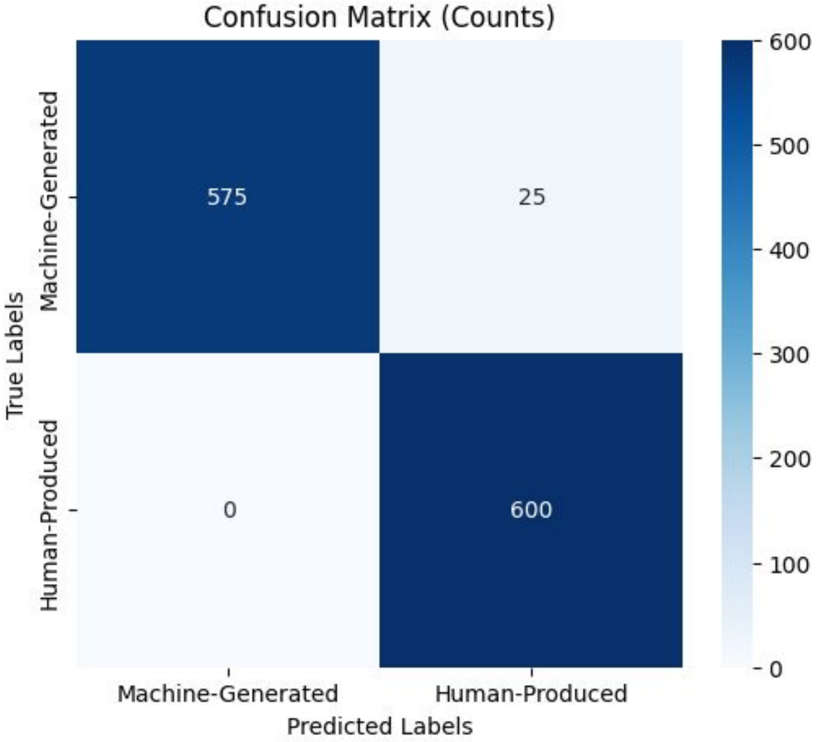
Recall: 1.0000



arxiv_davinci_test:

Accuracy: 0.9792
Precision: 1.0000
Recall: 0.9583



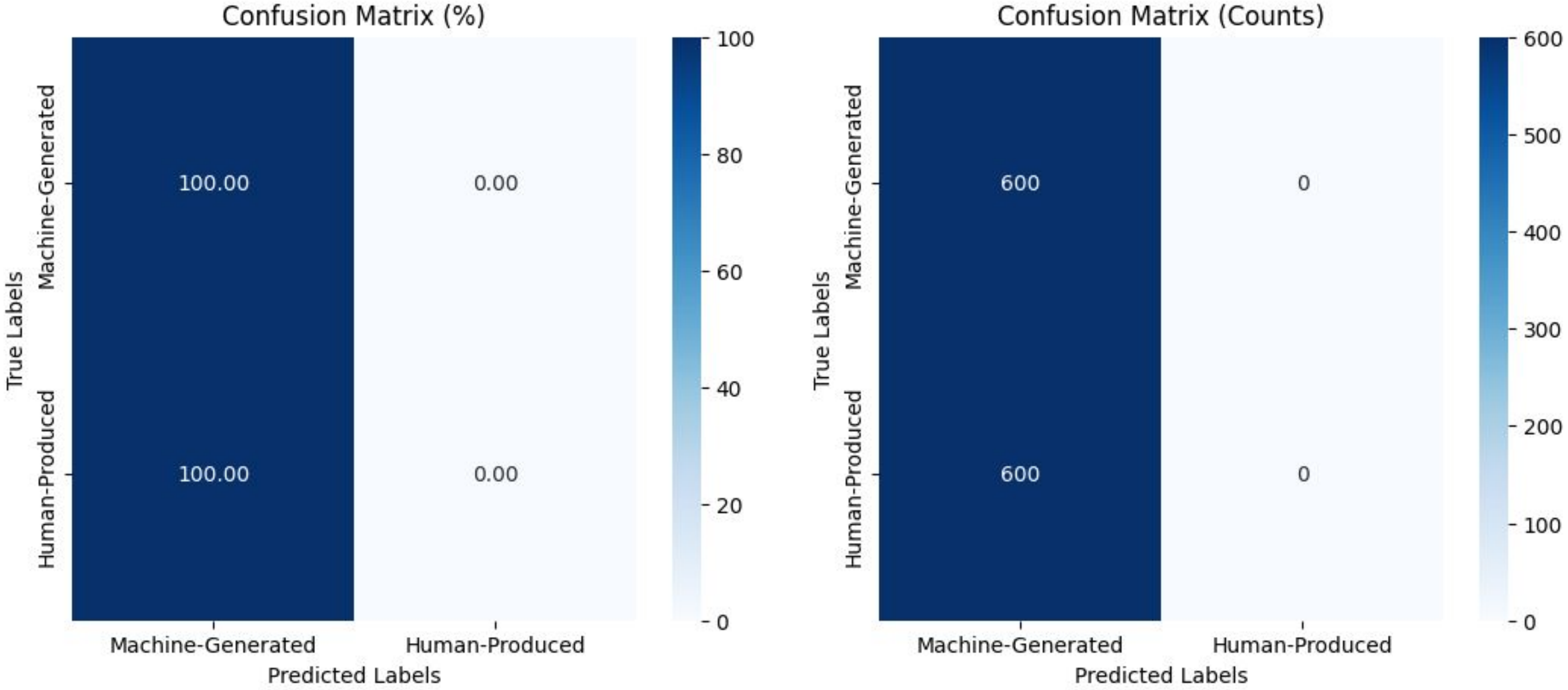


Bloomz-560m-academic-detector: Fine-tuned using arxiv_bloomz

TESTING WITH \N REMOVED

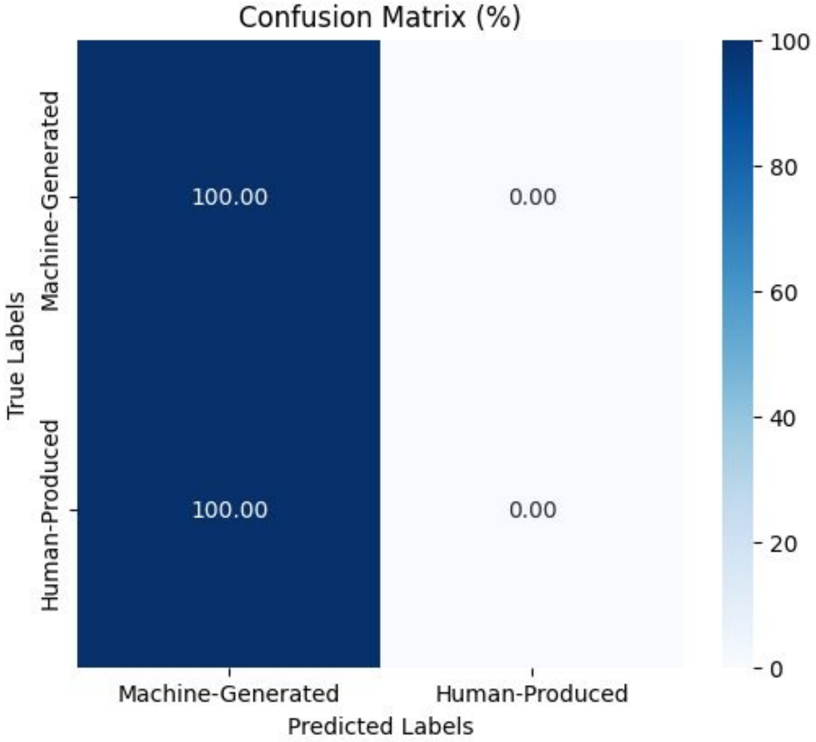
arxiv_bloomz_test:

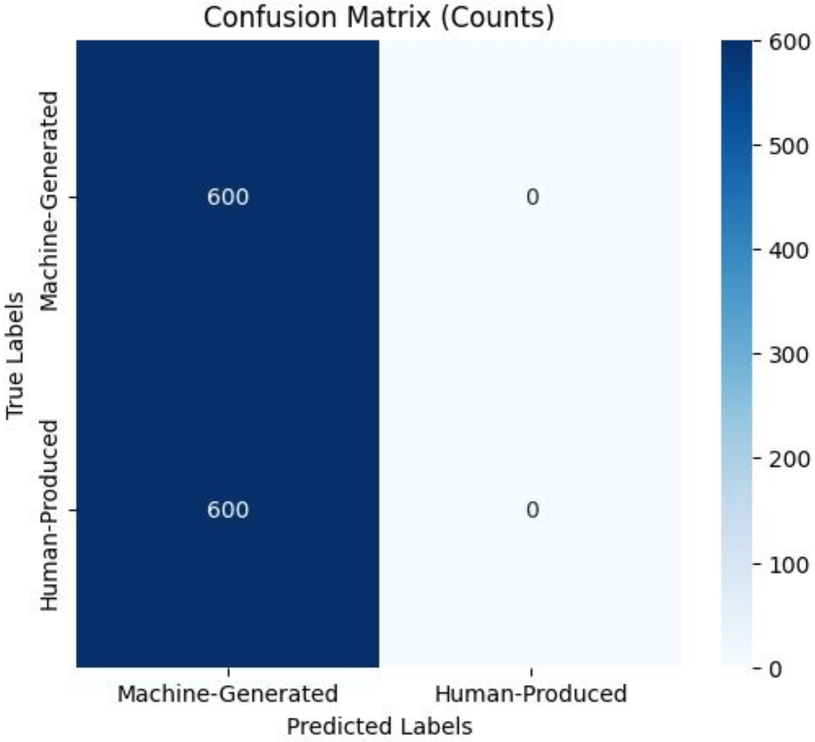
Accuracy: 0.5000
Precision: 0.5000
Recall: 1.0000



arxiv_chatGPT_test:

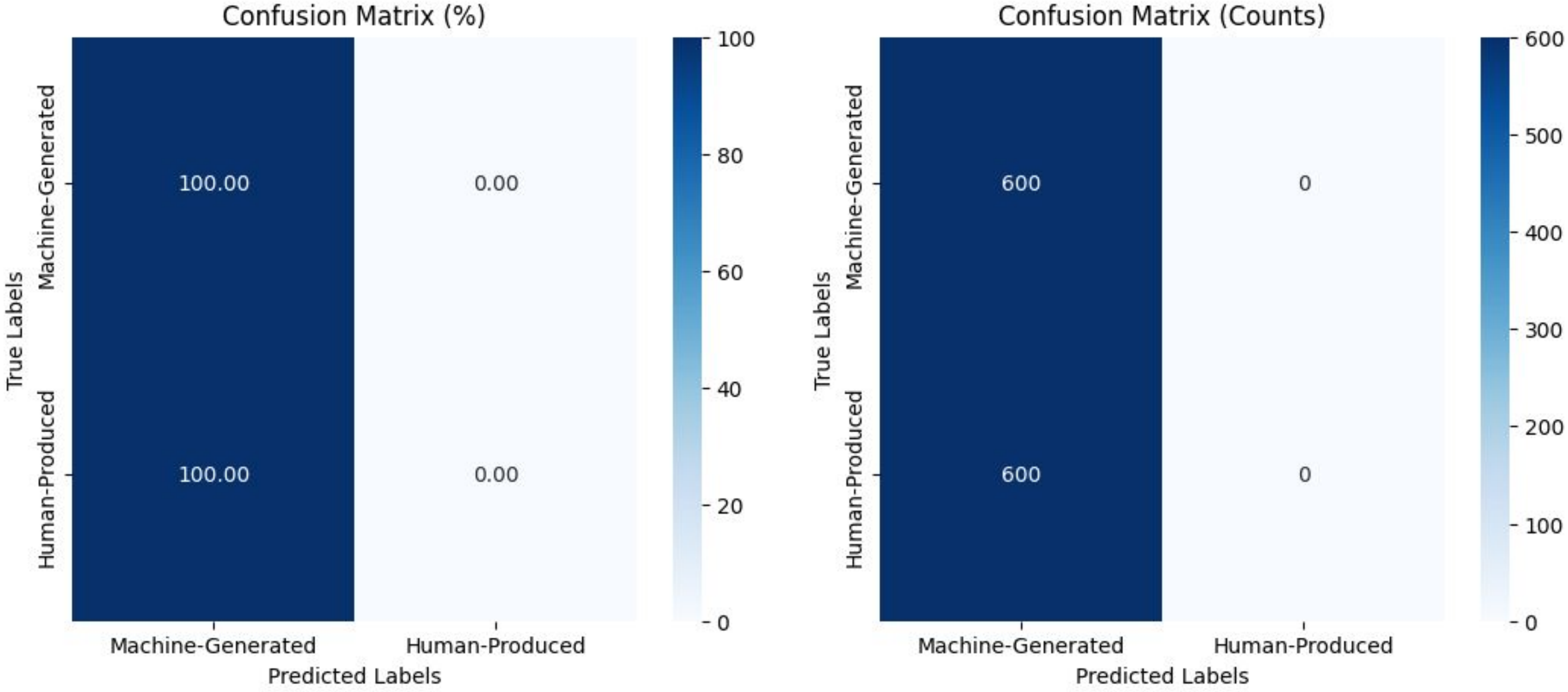
Accuracy: 0.5000
Precision: 0.5000
Recall: 1.0000





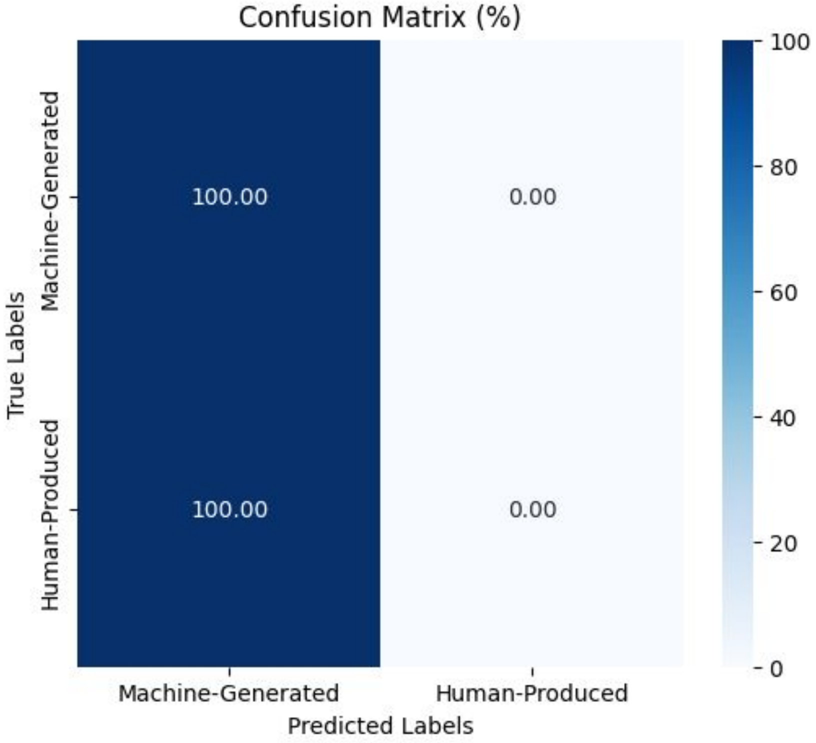
arxiv_cohere_test:

Accuracy: 0.5000
Precision: 0.5000
Recall: 1.0000

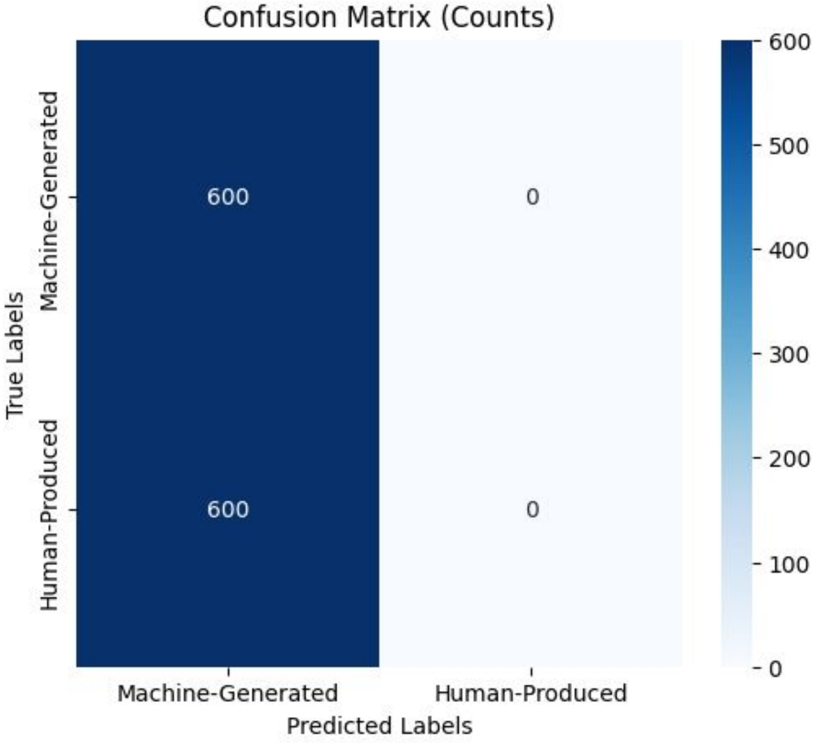


arxiv_davinci_test:

Accuracy: 0.5000
Precision: 0.5000
Recall: 1.0000



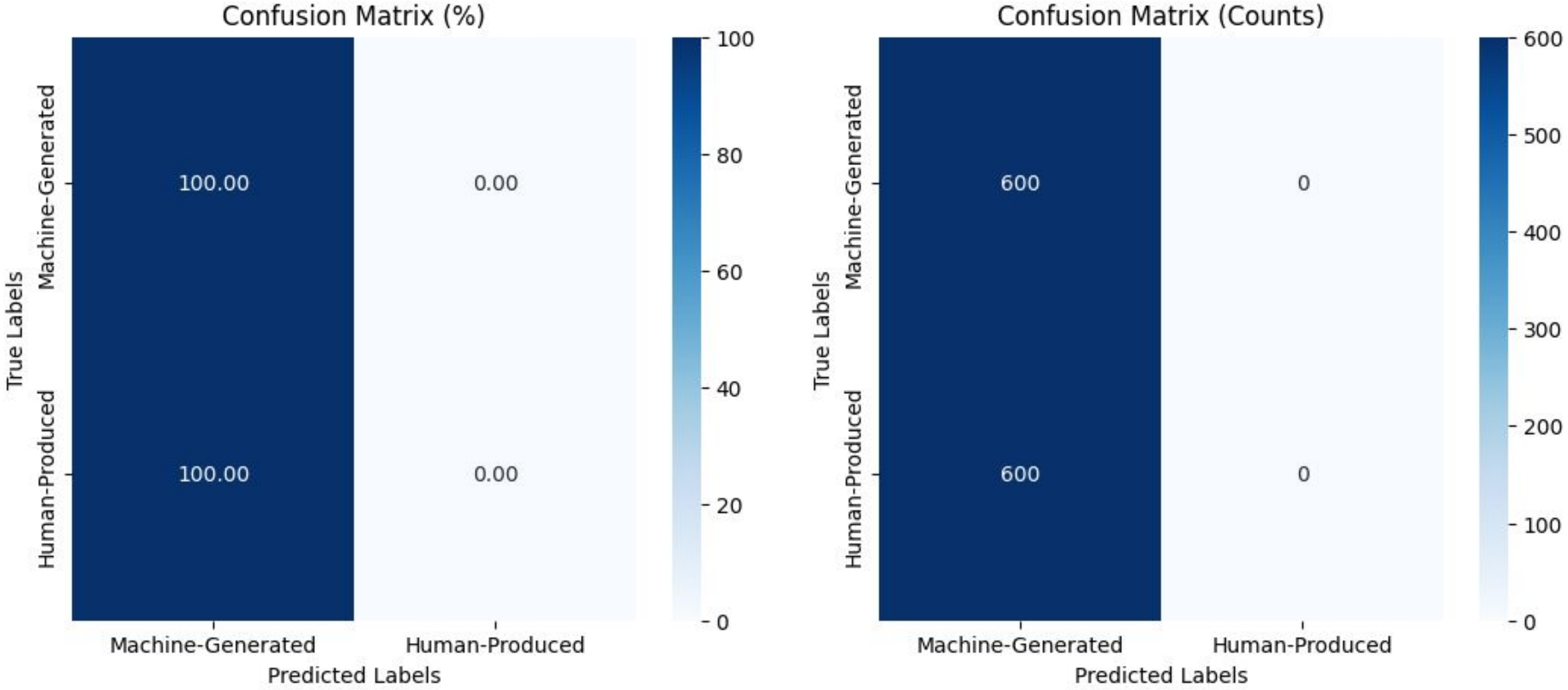
Confusion Matrix (%)



Confusion Matrix (Counts)

arxiv_flant5_test:

Accuracy: 0.5000
Precision: 0.5000
Recall: 1.0000

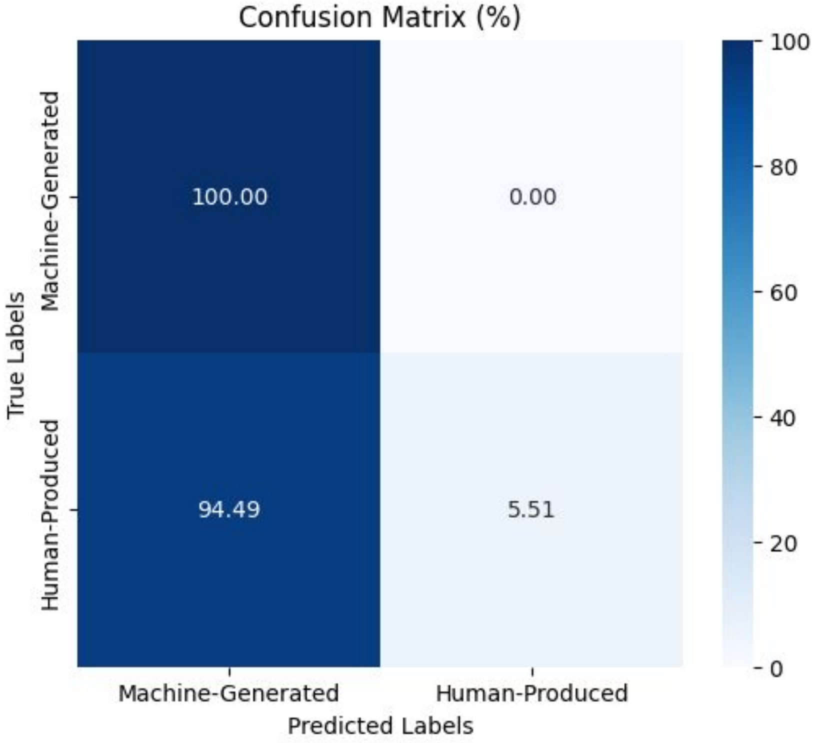


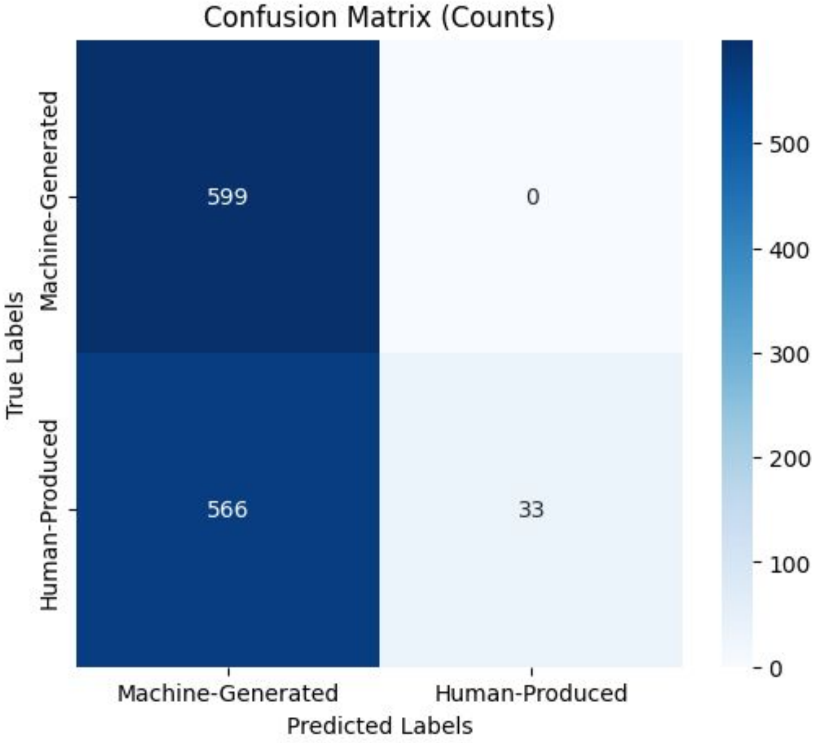
Bloomz-560m-academic-detector: Fine-tuned using arxiv_bloomz

TESTING WITH M4 WIKI

wikipedia_chatgpt_test:

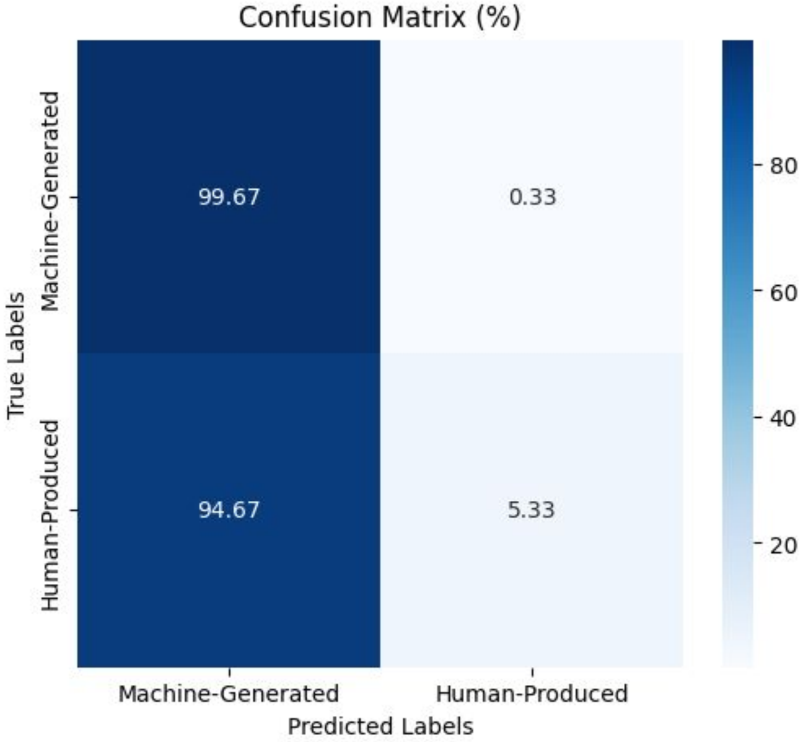
Accuracy: 0.5275
Precision: 0.5142
Recall: 1.0000

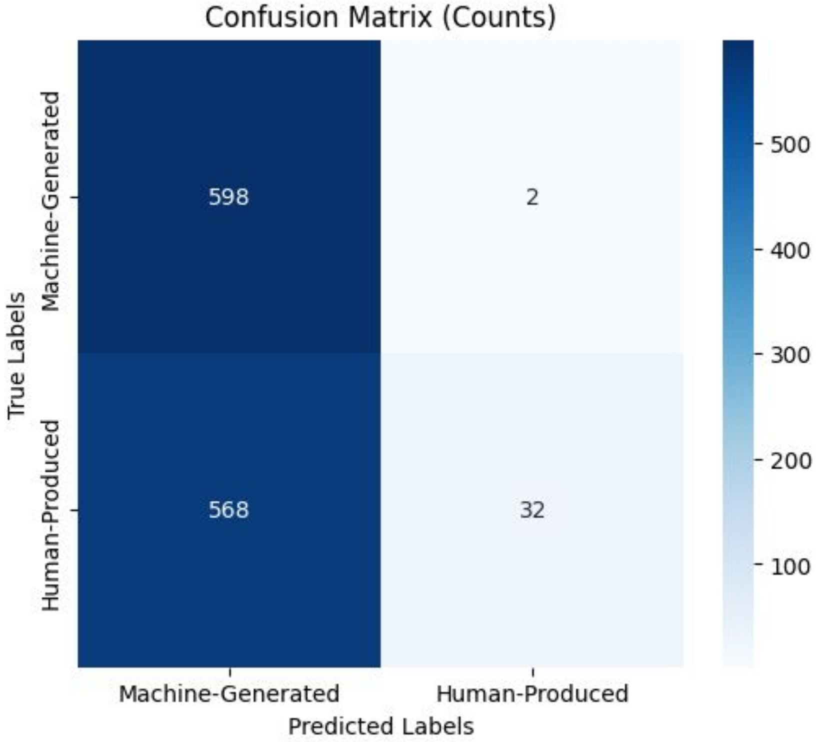




wikipedia_bloomz_test:

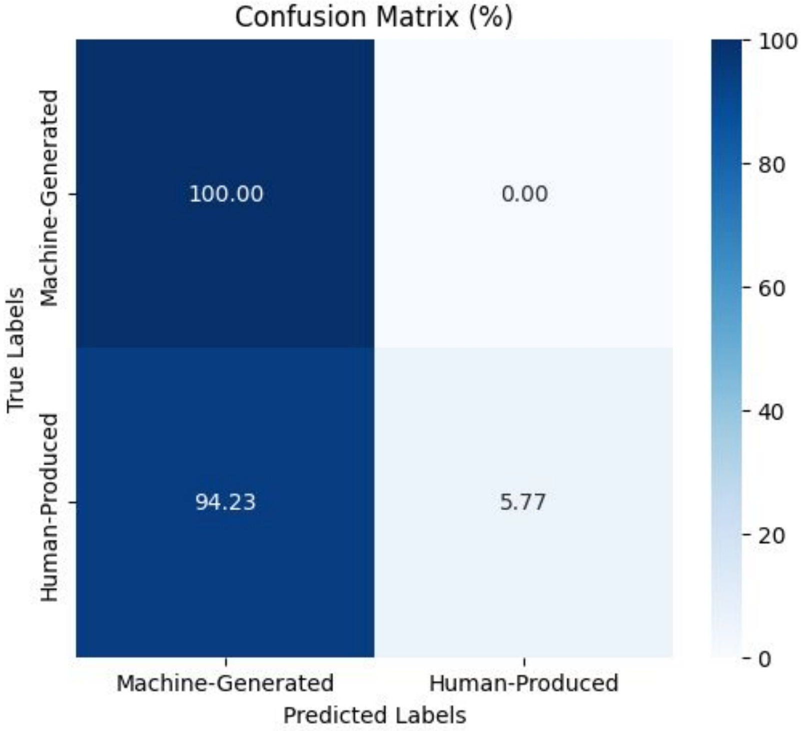
Accuracy: 0.5250
Precision: 0.5129
Recall: 0.9967

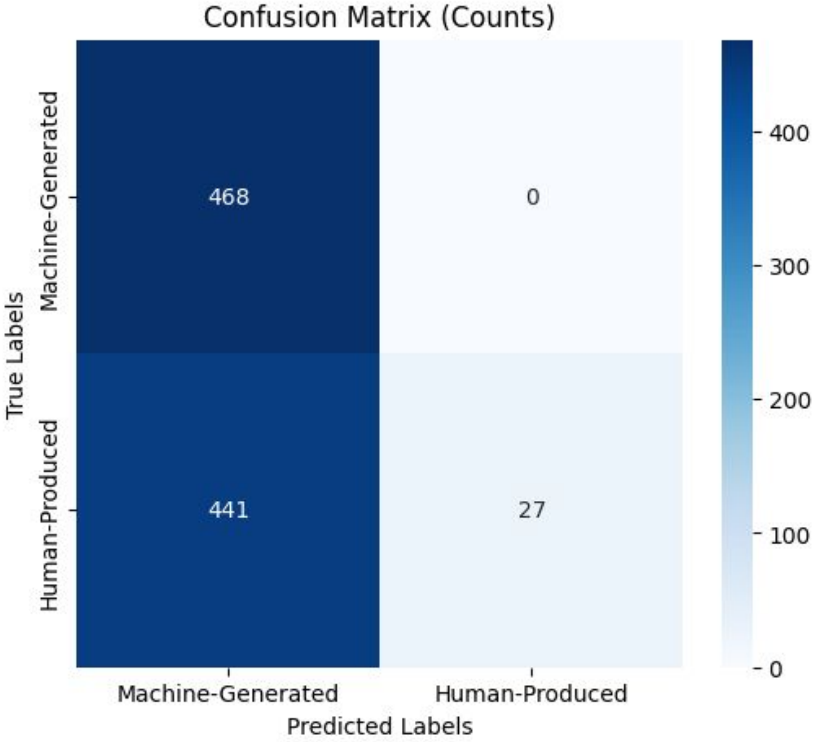




wikipedia_cohere_test:

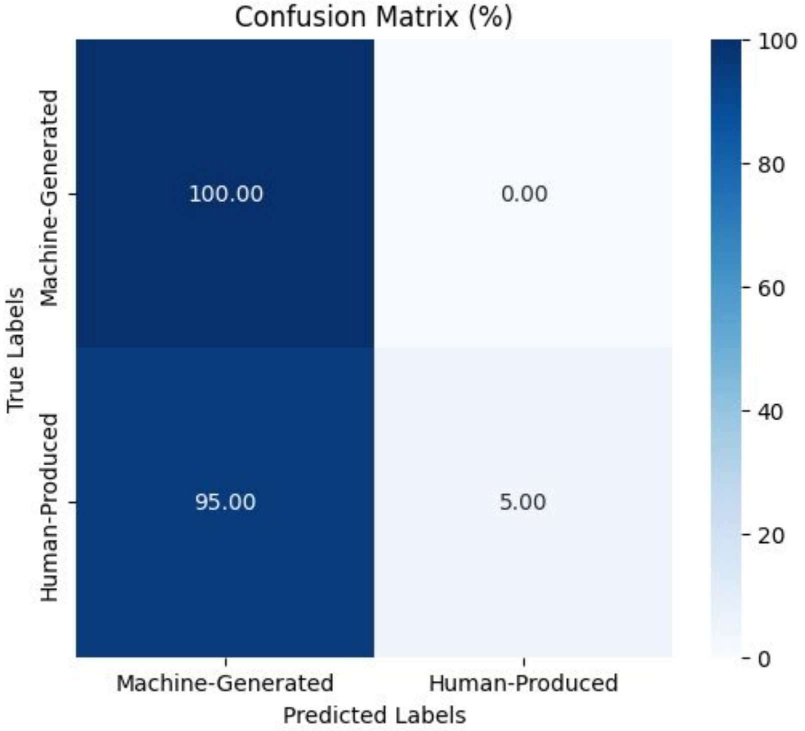
Accuracy: 0.5288
Precision: 0.5149
Recall: 1.0000

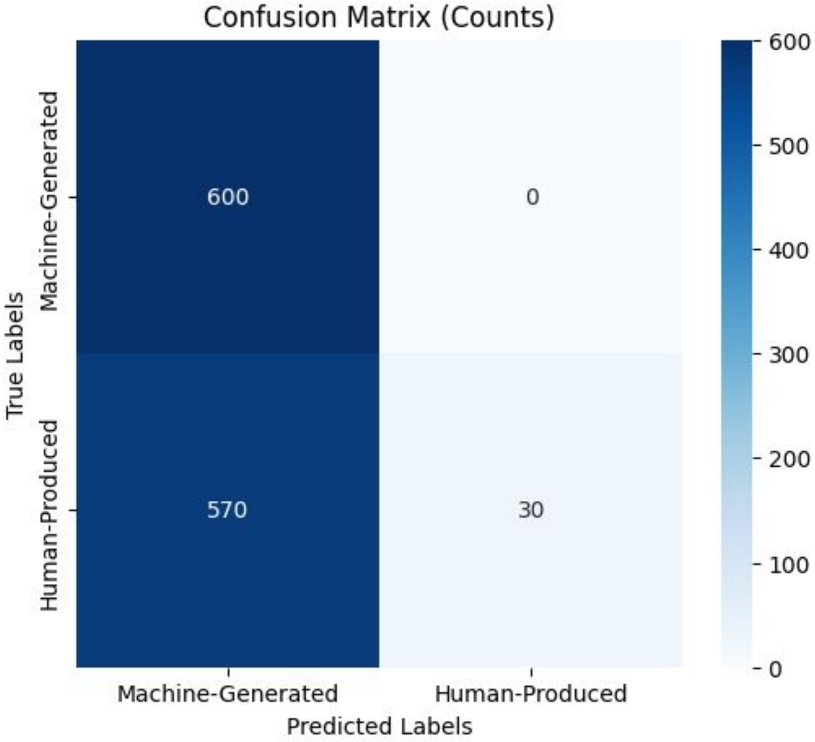




wikipedia_davinci_test:

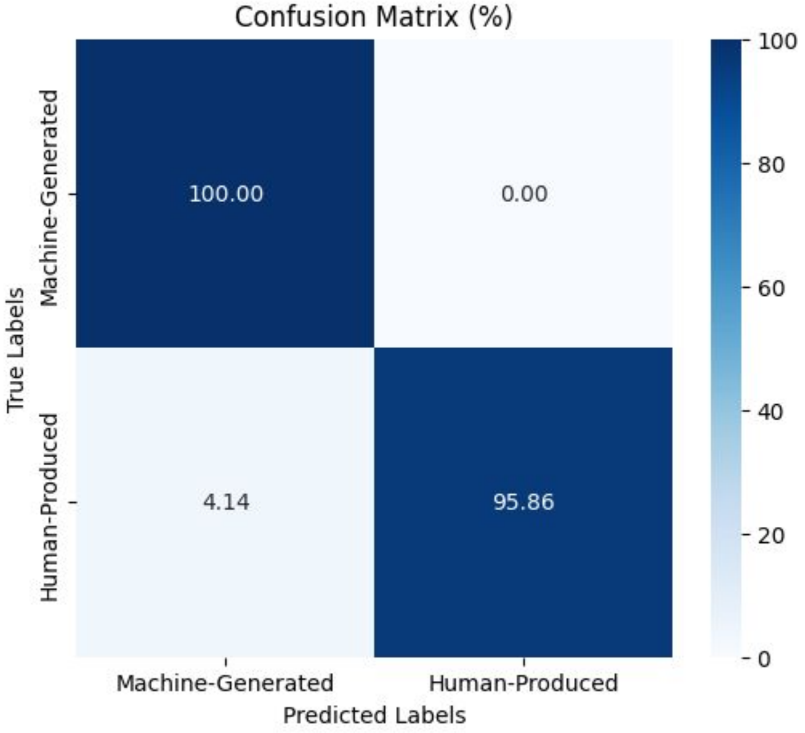
Accuracy: 0.5250
Precision: 0.5128
Recall: 1.0000

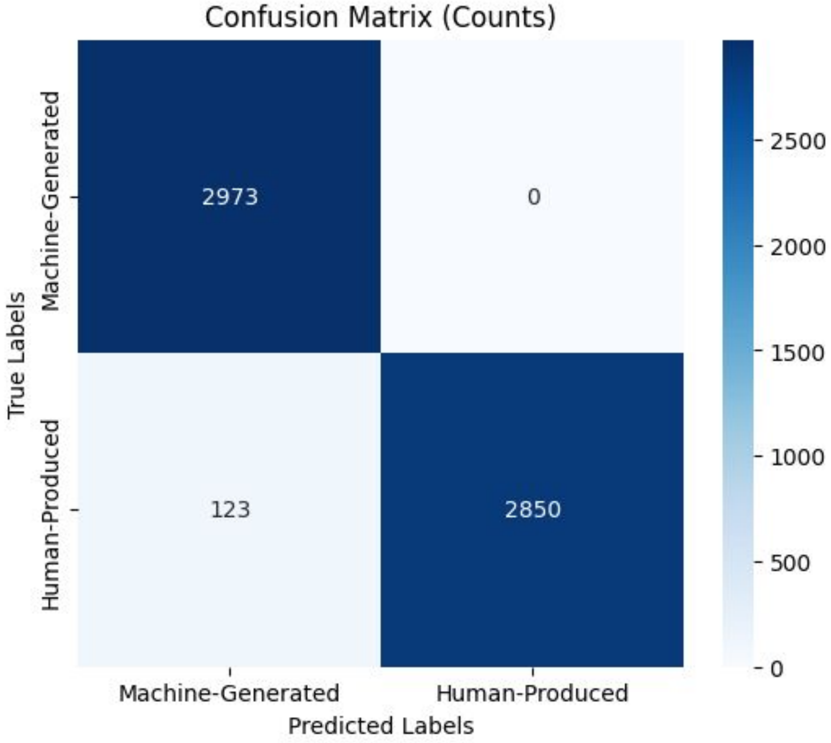




Testing with ML dataset:

Accuracy: 0.9793
Precision: 0.9603
Recall: 1.0000


Confusion Matrix (%)


Confusion Matrix (Counts)

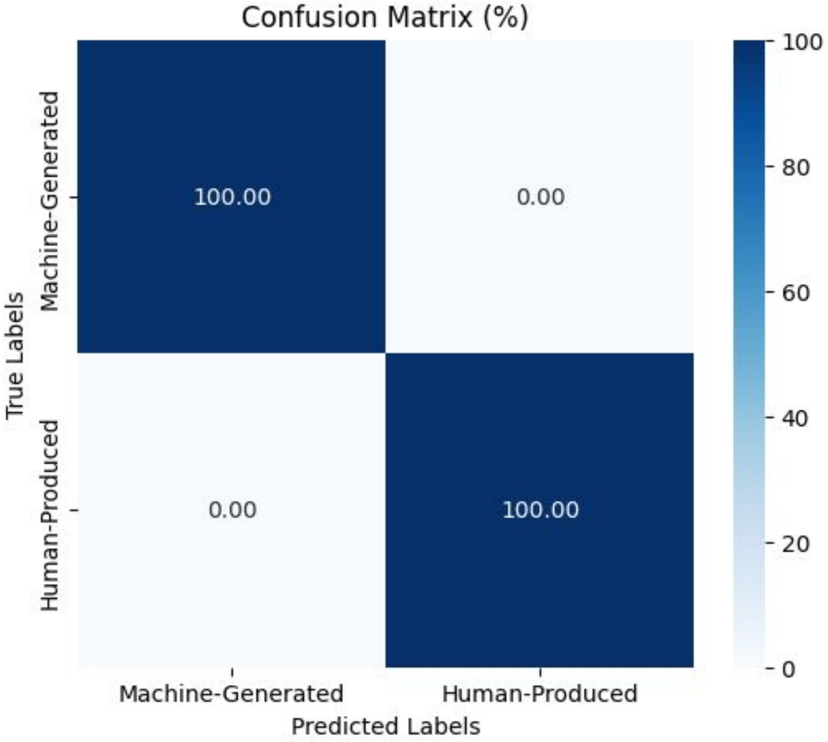
FINETUNING & TESTING WITH \N REMOVED DATASETS

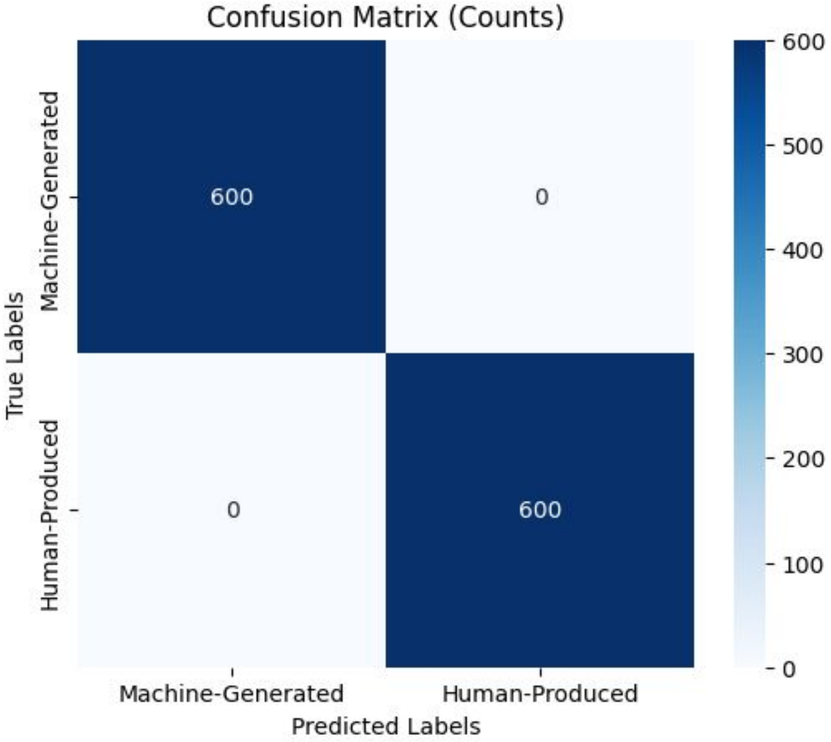
Bloomz-560m-academic-detector:
Fine-tuned **arxiv_bloomz_\n_removed**

Test using \n_removed_(chatGPT,cohere,flant5,davinci,chatGPT_para,bloomz_para)_test_set

Testing with arxiv_bloomz_\n_removed held-out test_dataset:

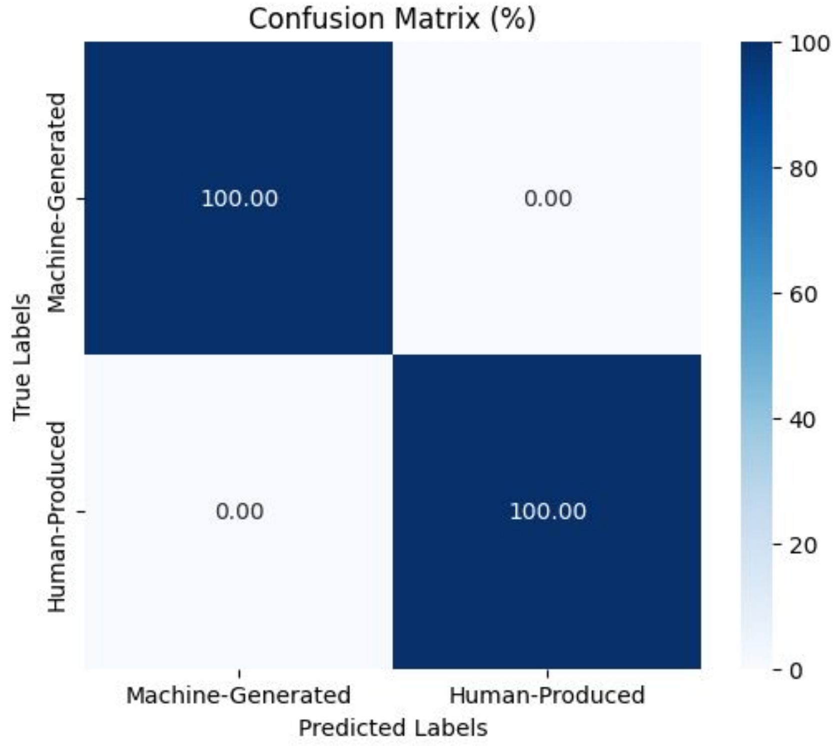
Accuracy: 1.0000
Precision: 1.0000
Recall: 1.0000

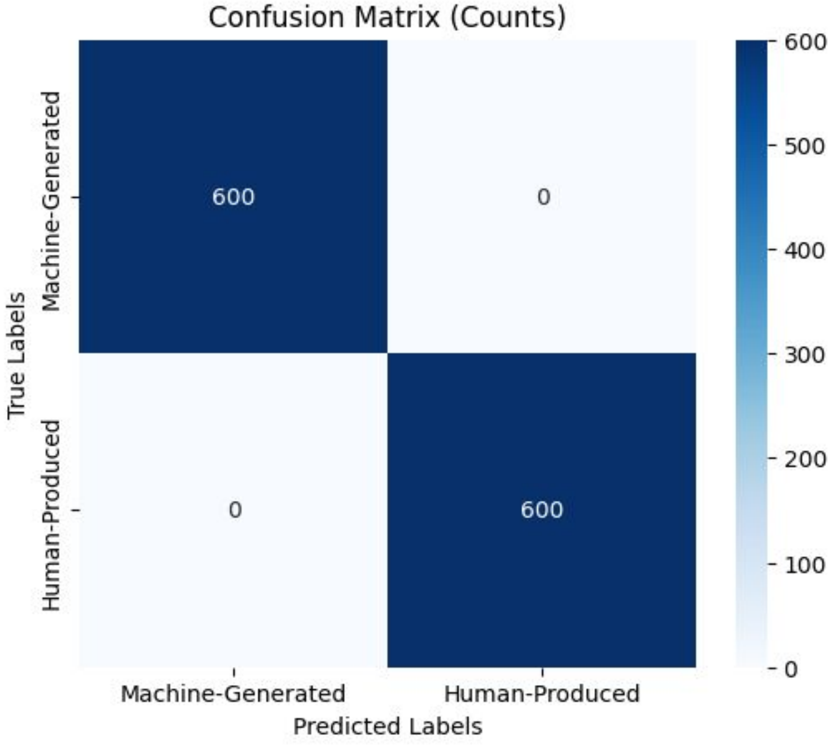




Testing with arxiv_\n_removed test_data from ChatGPT:

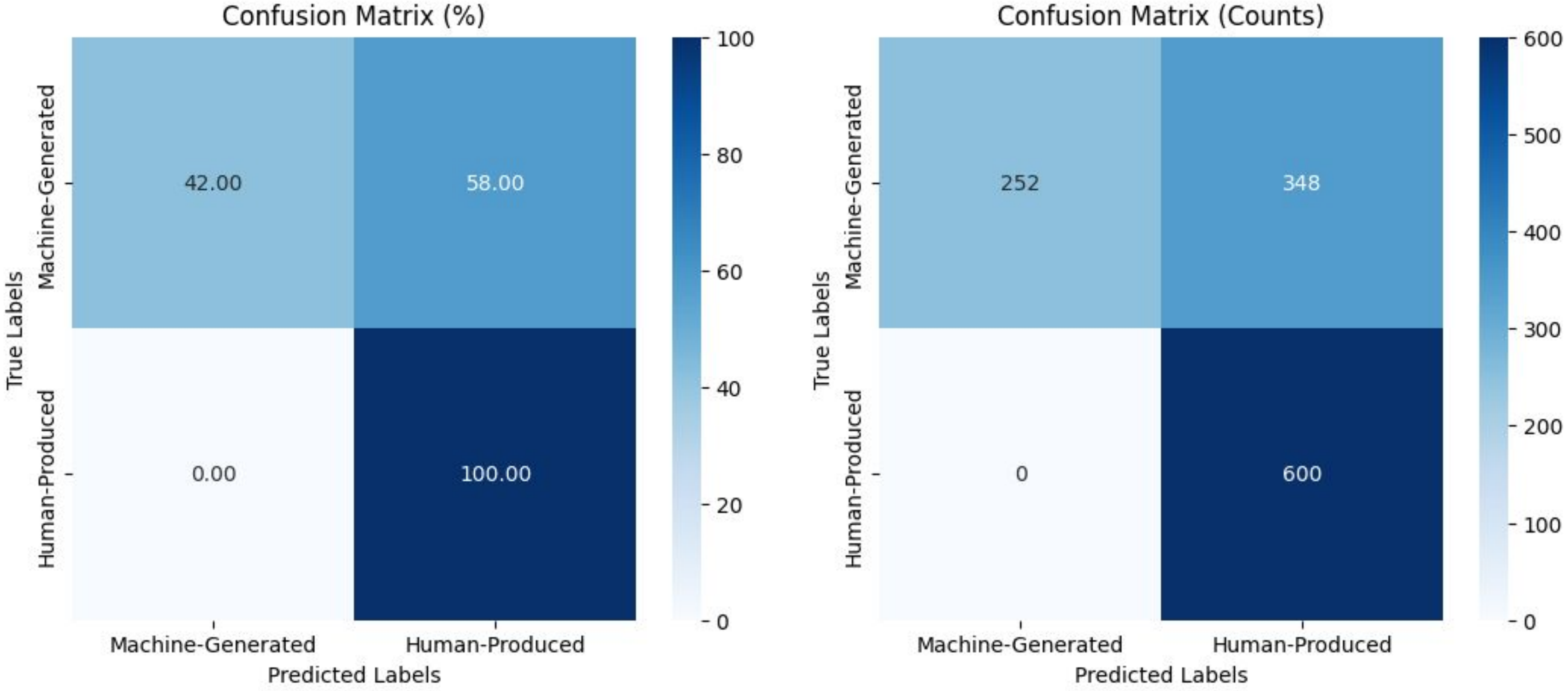
Accuracy: 1.0000
Precision: 1.0000
Recall: 1.0000





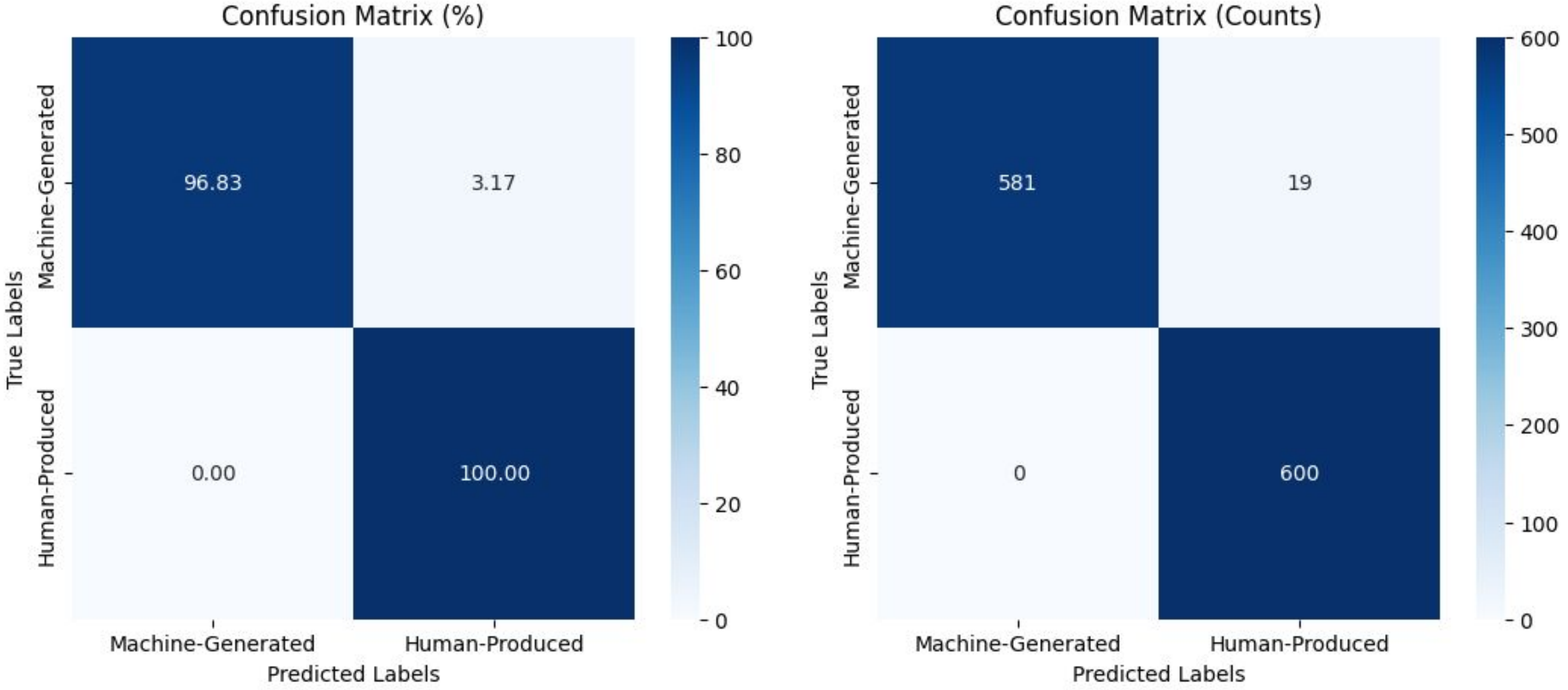
Testing with arxiv_\n_removed test_data from Cohere:

Accuracy: 0.7100
Precision: 1.0000
Recall: 0.4200



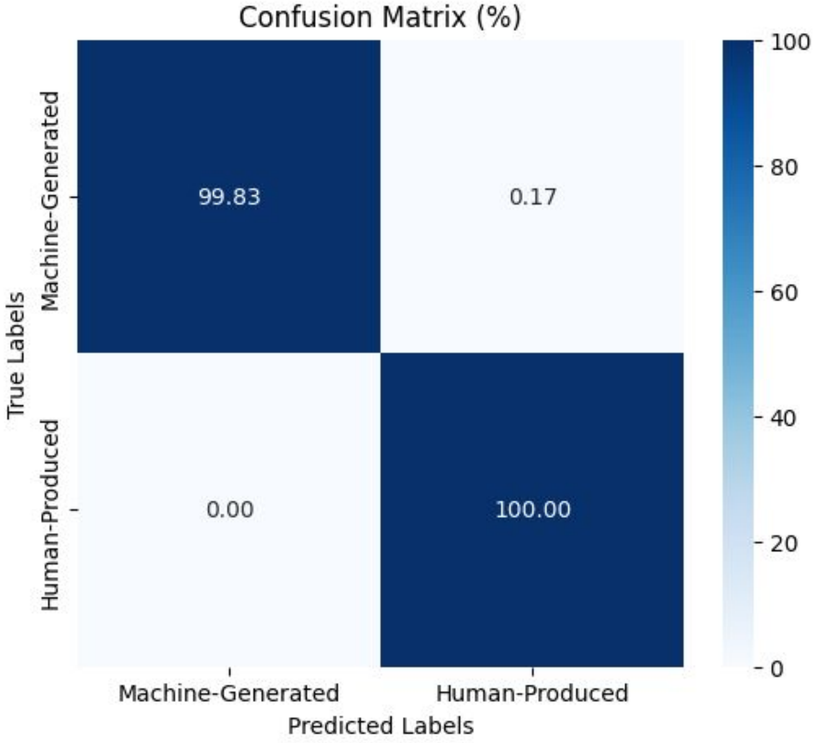
Testing with arxiv_\n_removed test_data from Davinci:

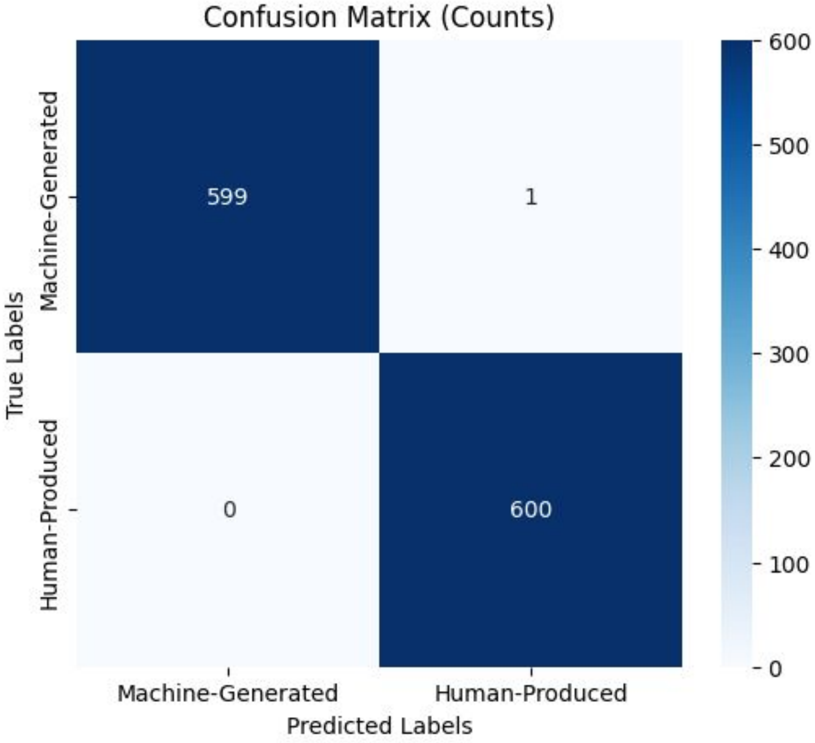
Accuracy: 0.9842
Precision: 1.0000
Recall: 0.9683



Testing with arxiv_\n_removed test_data from Flant5:

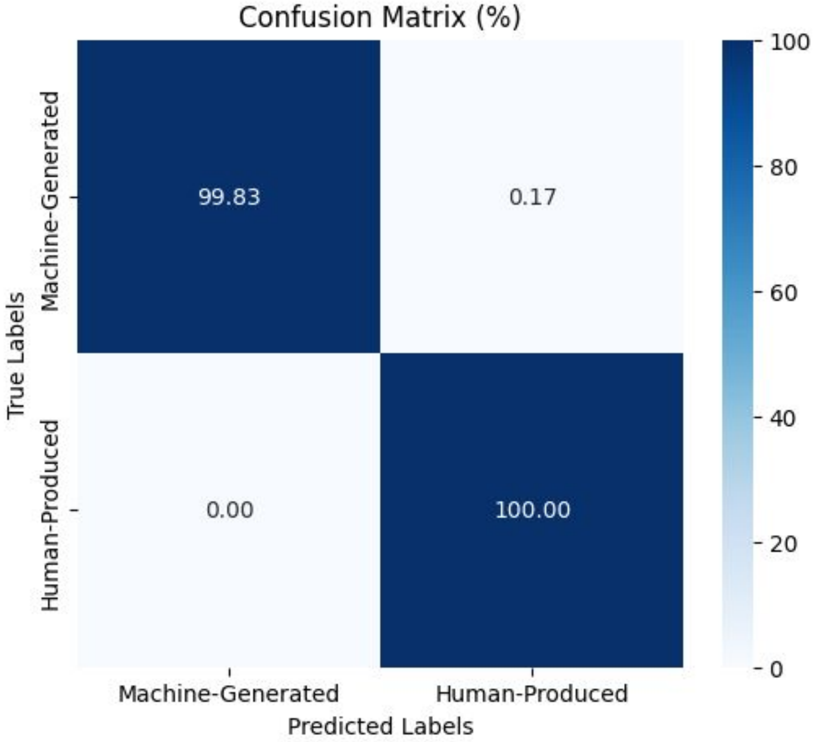
Accuracy: 0.9992
Precision: 1.0000
Recall: 0.9983

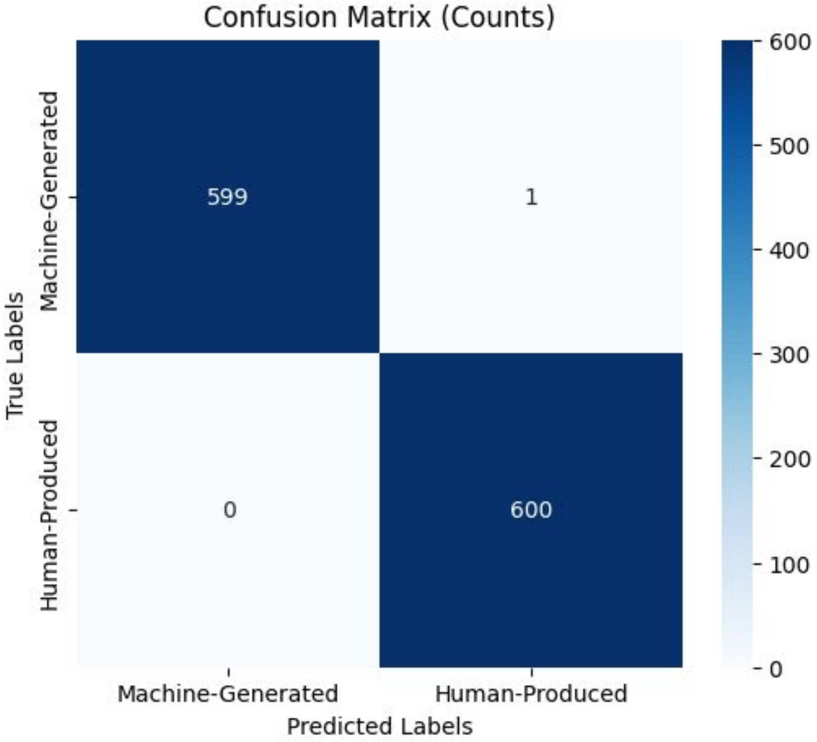




Testing with arxiv_\n_removed test_data from Flant5:

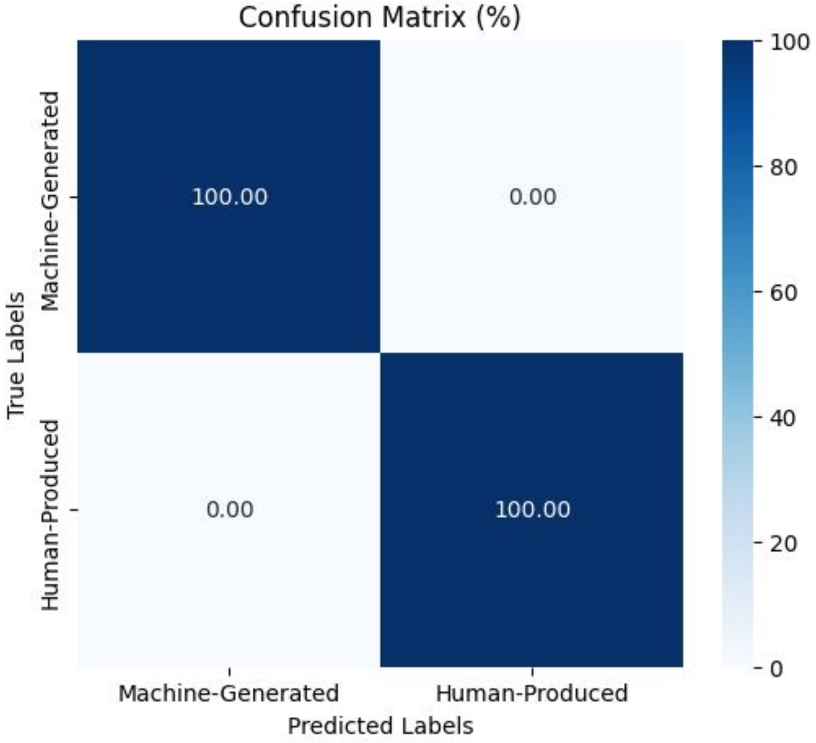
Accuracy: 0.9992
Precision: 1.0000
Recall: 0.9983



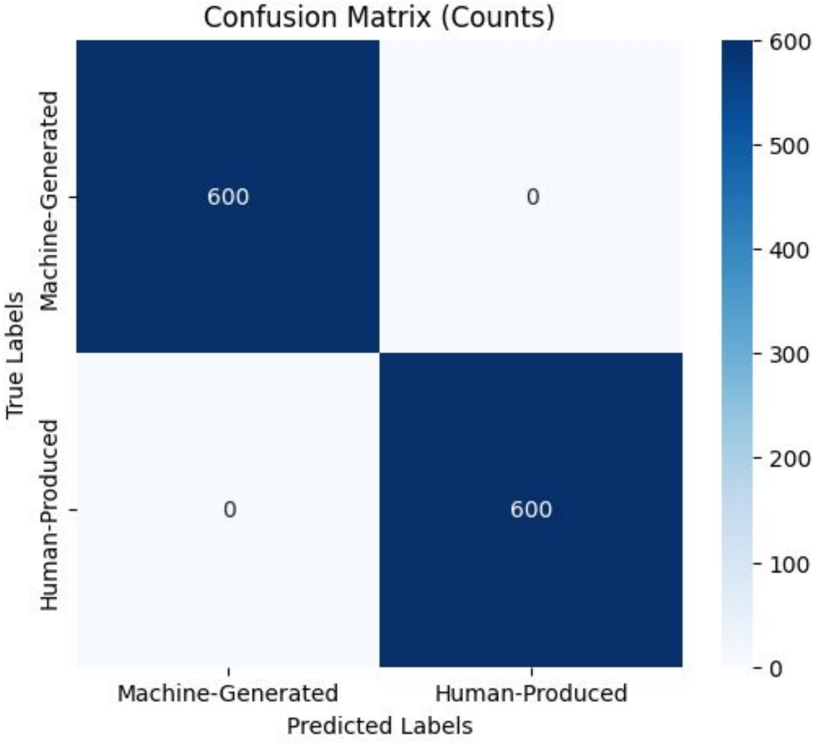


Testing with arxiv_\n_removed test_data from bloomz_paraphrased:

Accuracy: 1.0000
Precision: 1.0000
Recall: 1.0000



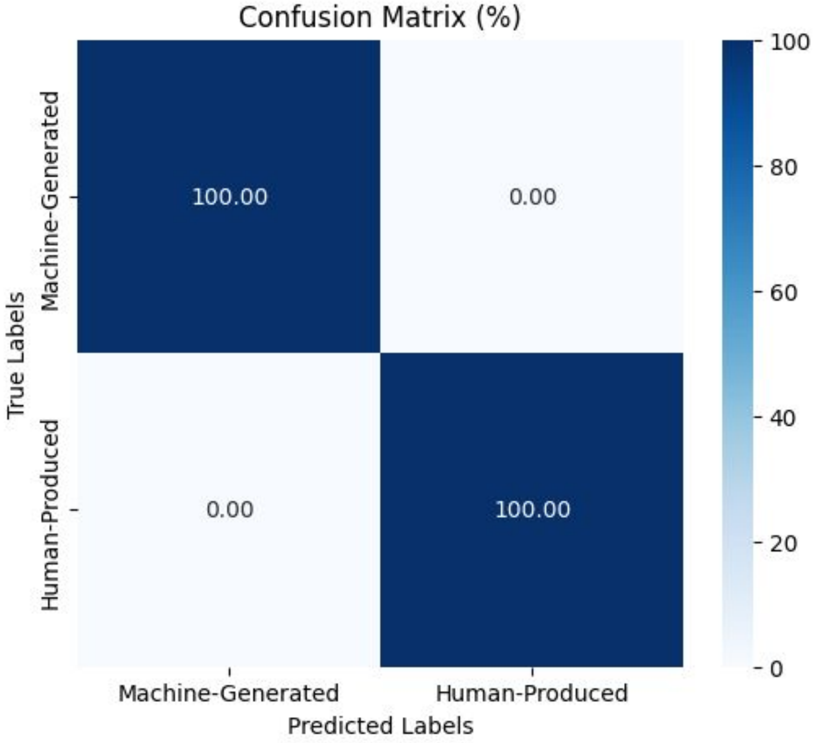
Confusion Matrix (%)

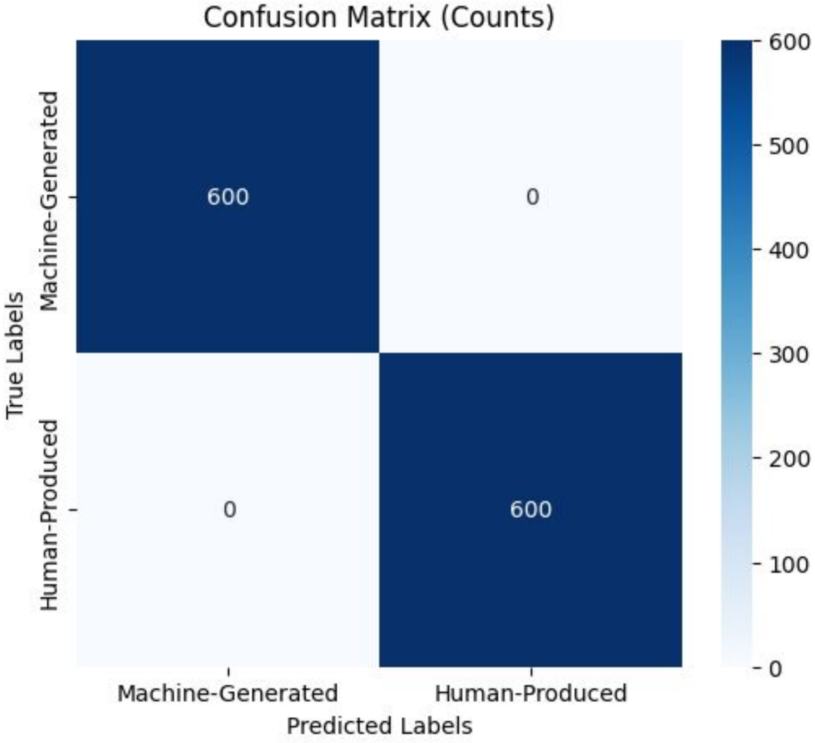


Confusion Matrix (Counts)

Testing with arxiv_\n_removed test_data from chatGPT_paraphrased:

Accuracy: 1.0000
Precision: 1.0000
Recall: 1.0000



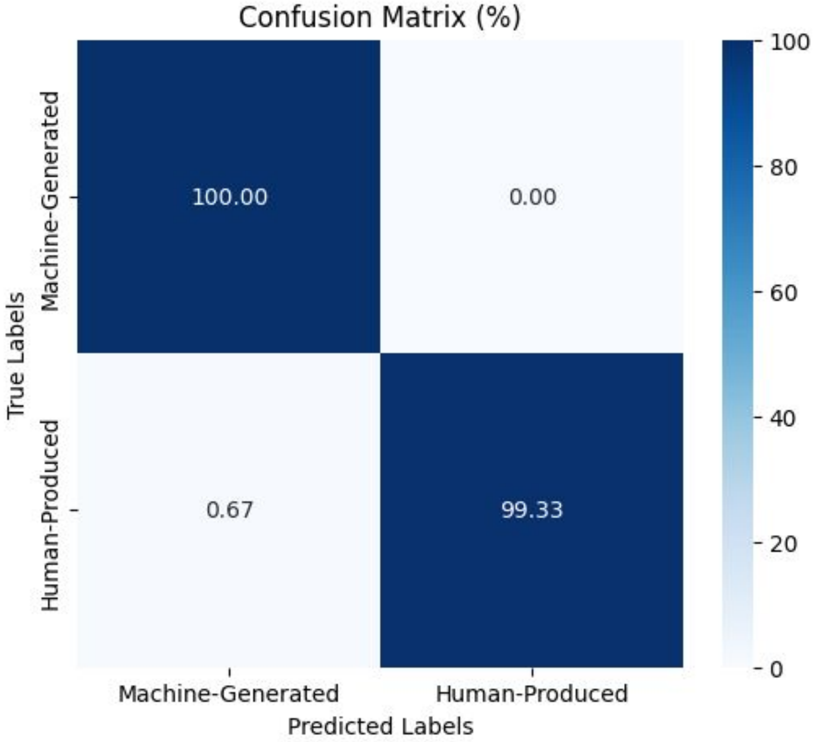


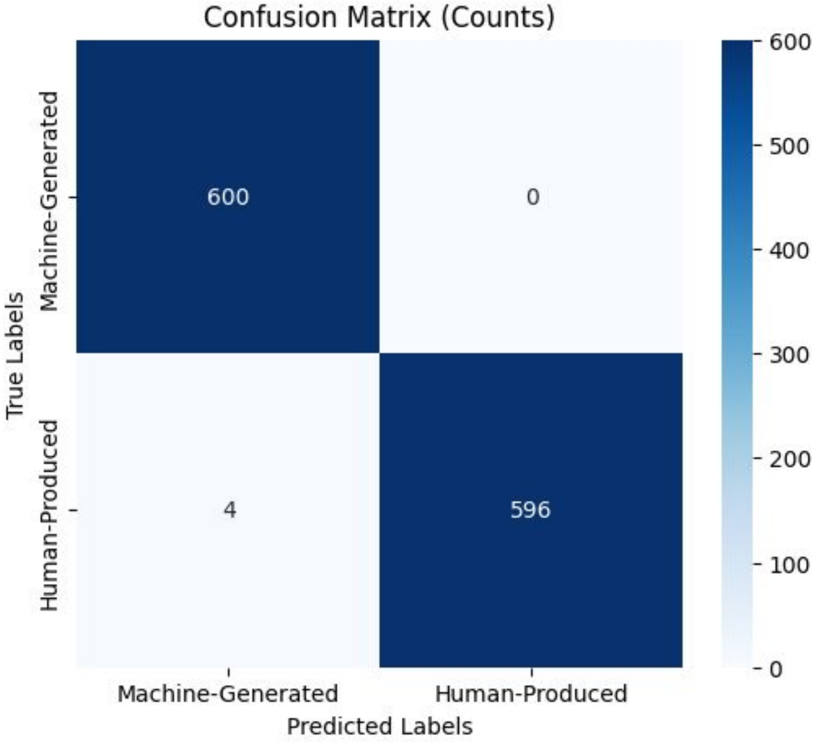
Bloomz-560m-academic-detector:
Fine-tuned **arxiv_bloomz_\n_removed + wikipedia_bloomz_\n_removed**

Test using \n_removed_wiki_(chatGPT,cohere,davinci)_test_set

Testing with wikipedia_bloomz_\n_removed held-out test_dataset:

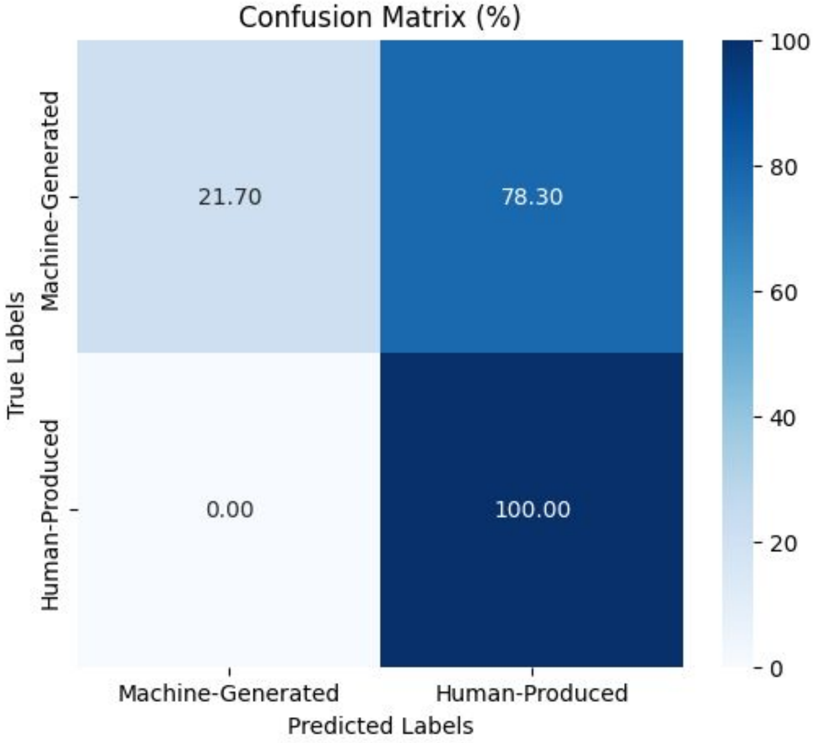
Accuracy: 0.9967
Precision: 0.9934
Recall: 1.0000

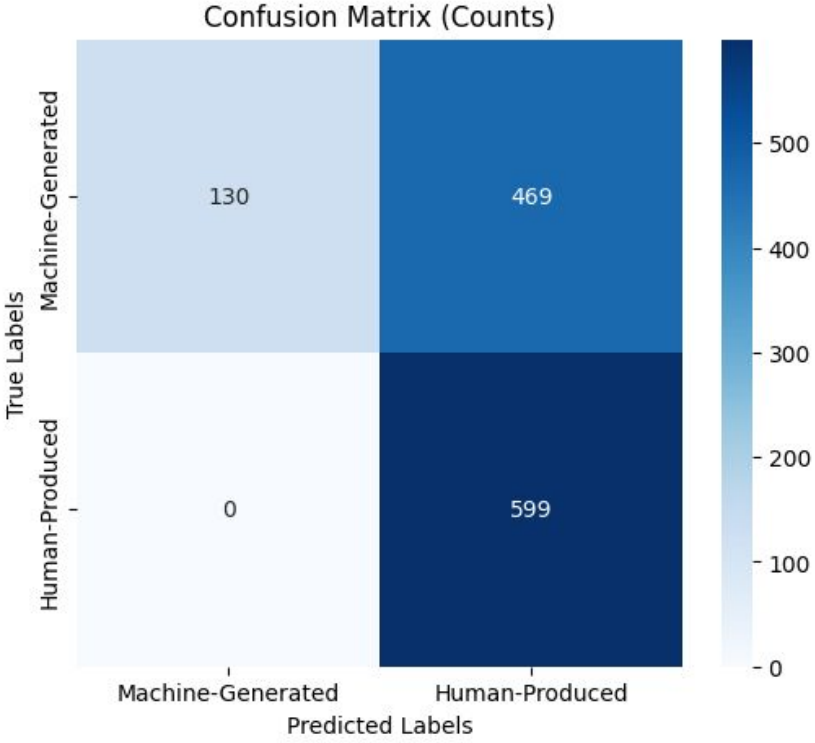




Testing with wikipedia_chatgpt_test_\n_removed test_dataset:

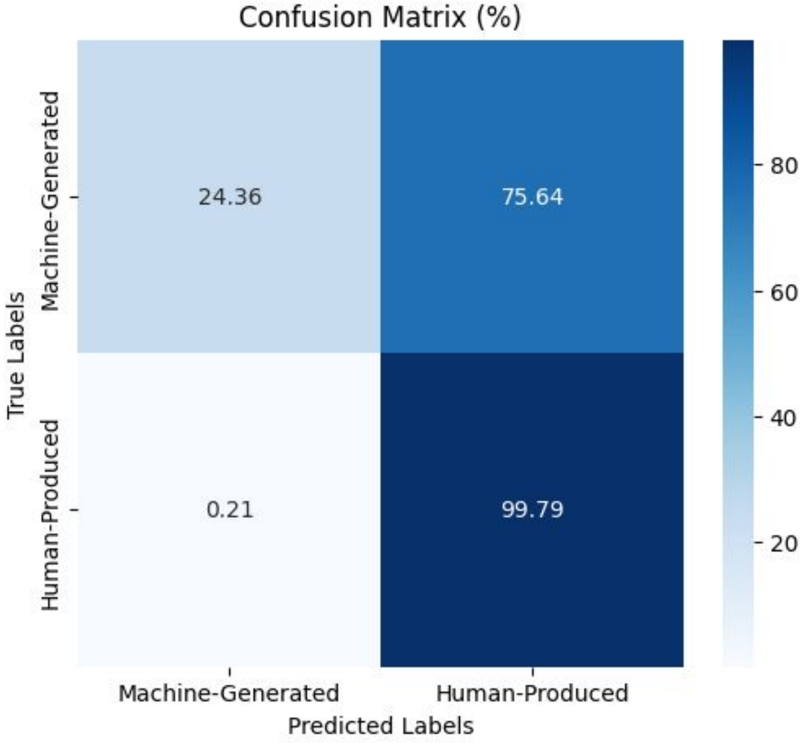
Accuracy: 0.6085
Precision: 1.0000
Recall: 0.2170

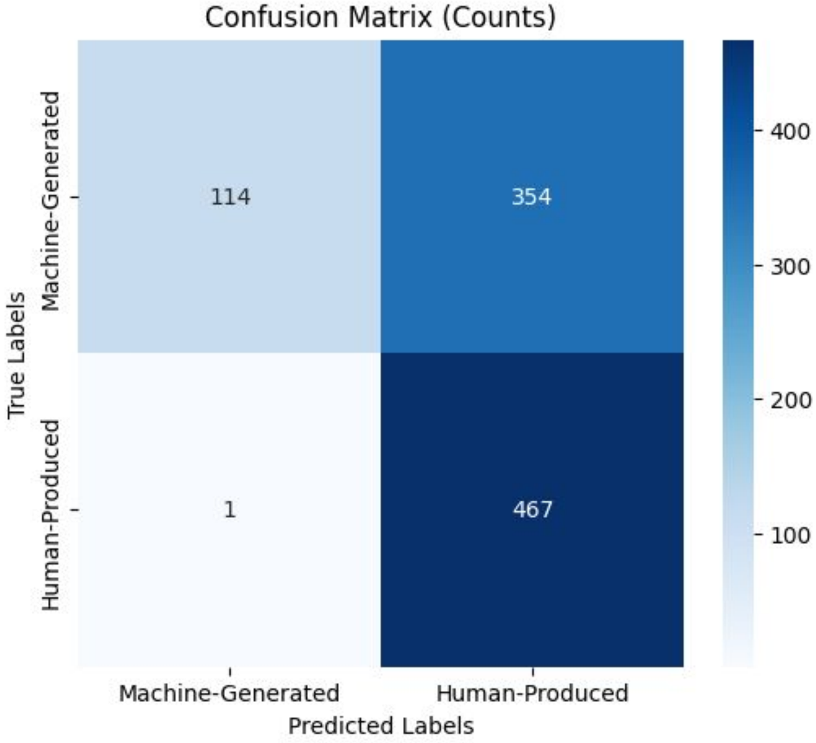




Testing with wikipedia_cohere_test_\n_removed:

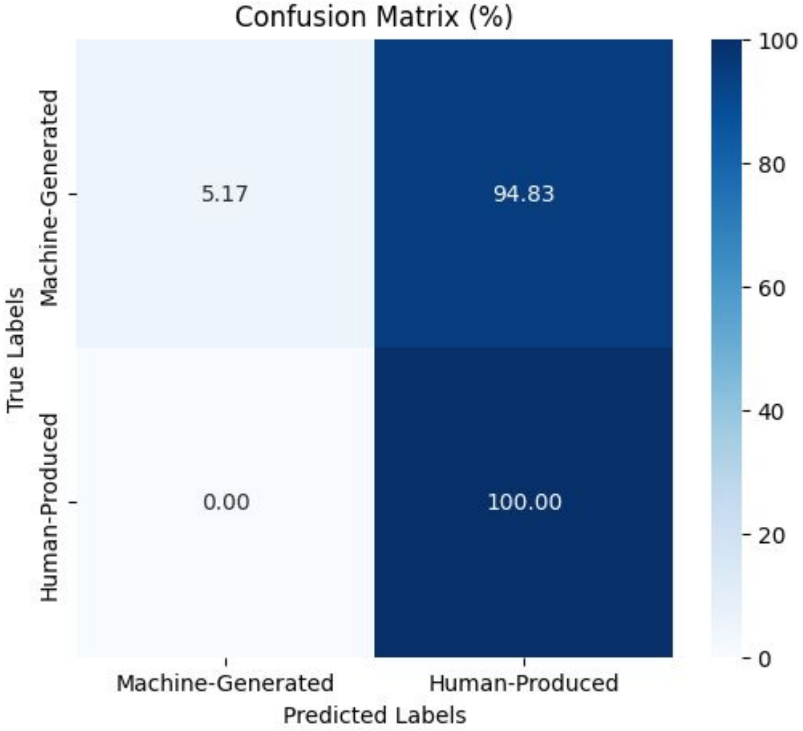
Accuracy: 0.6207
Precision: 0.9913
Recall: 0.2436

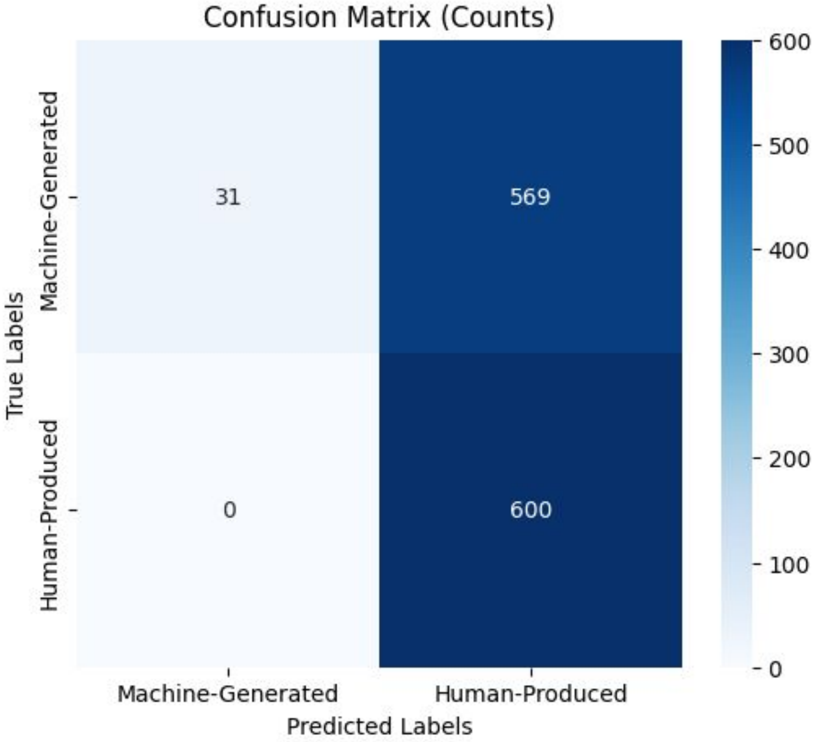




Testing with wikipedia_davinci_test_\n_removed:

Accuracy: 0.5258
Precision: 1.0000
Recall: 0.0517



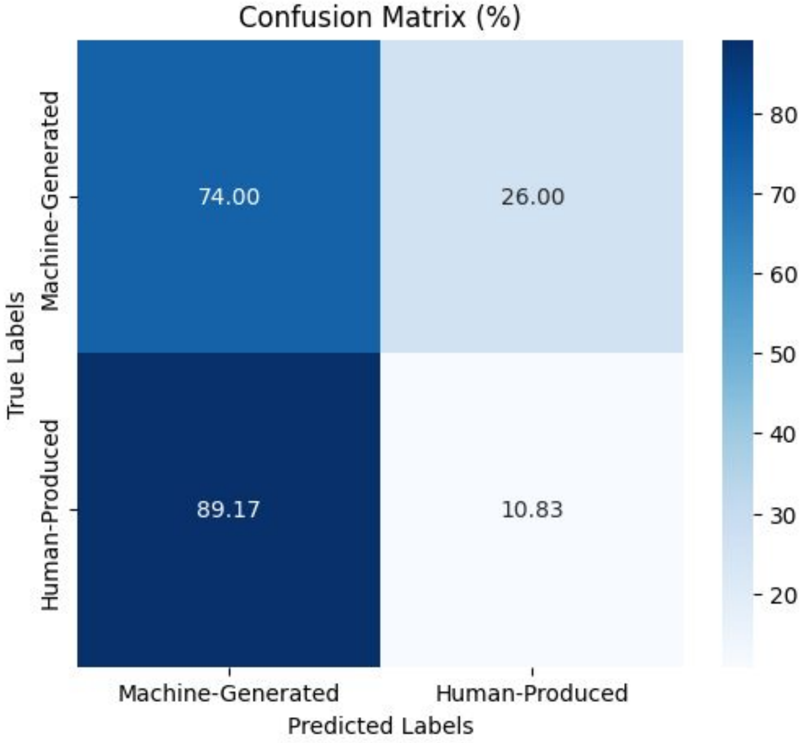


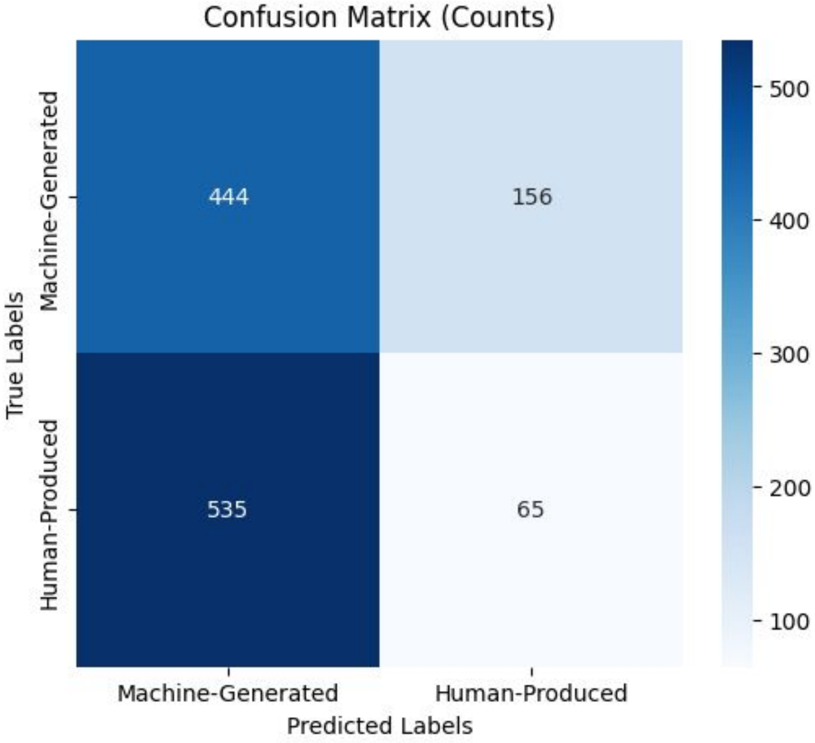
Bloomz-560m-academic-detector:
Fine-tuned **arxiv_bloomz_\n_removed + wikipedia_bloomz_\n_removed + ML_dataset_\n_removed**

Test using ML_data_test_set_\n_removed

Testing with ML_dataset_\n_removed held-out test_dataset:

Accuracy: 0.4242
Precision: 0.4535
Recall: 0.7400



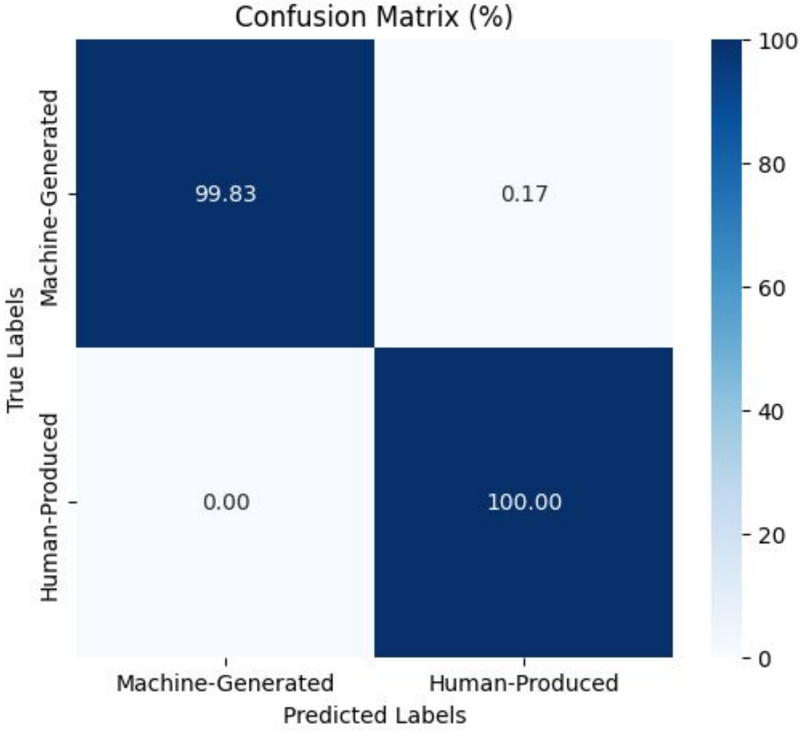


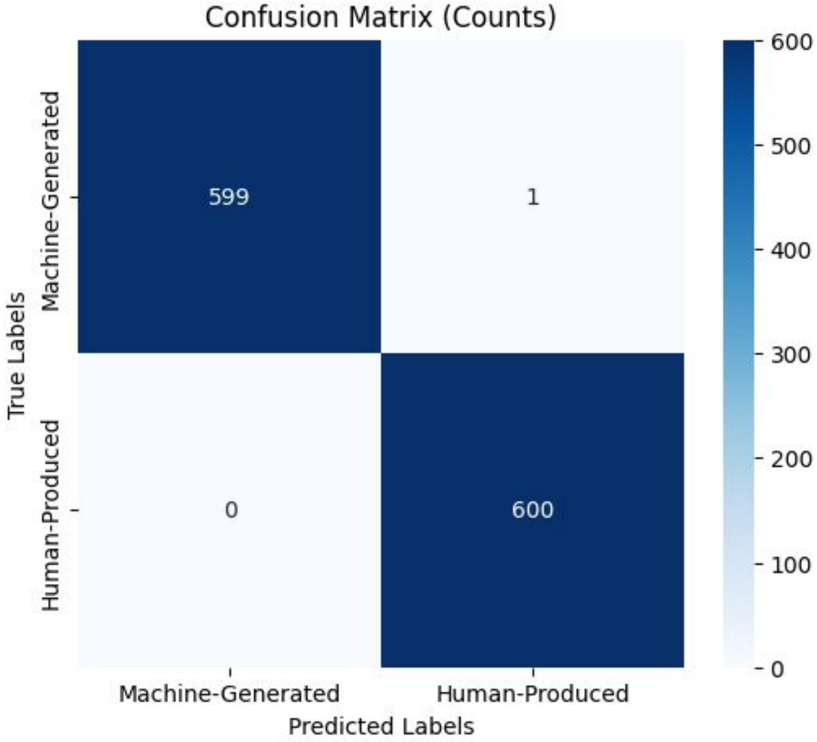
Bloomz-560m-academic-detector: **finetune using ML_dataset_\n_removed**

Test using ML_data_test_set_\n_removed

Testing with ML_dataset_\n_removed held-out test_dataset:

Accuracy: 0.9992
Precision: 1.0000
Recall: 0.9983



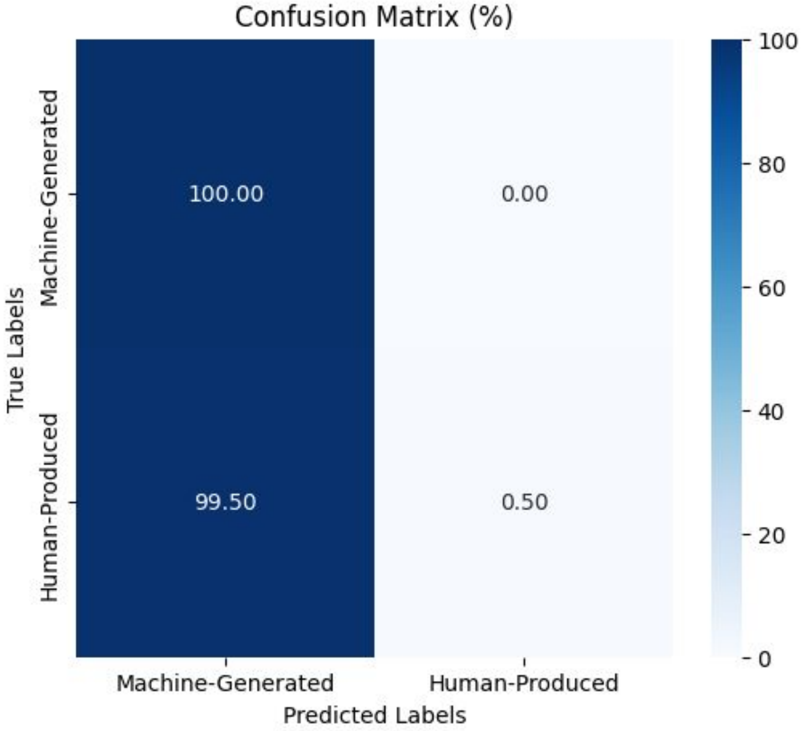


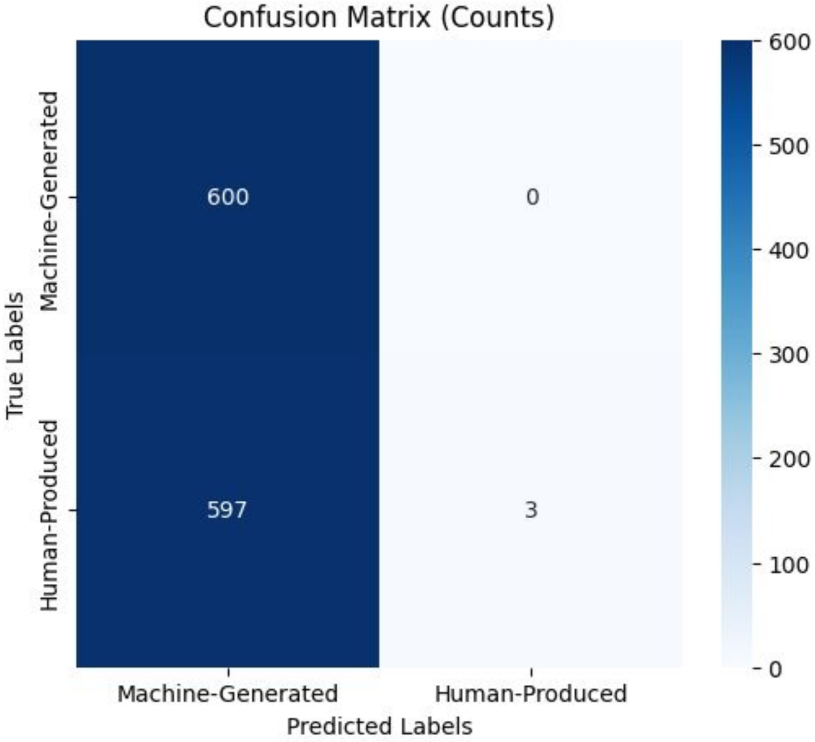
Bloomz-560m-academic-detector: **finetune using arxiv_bloomz_\n_removed**

Test using wiki_(chatGPT,cohere,davinci,bloomz)_test_set_with_\n_without_unicode

Testing with wikipedia_bloomz_test_with_\n_without_unicode

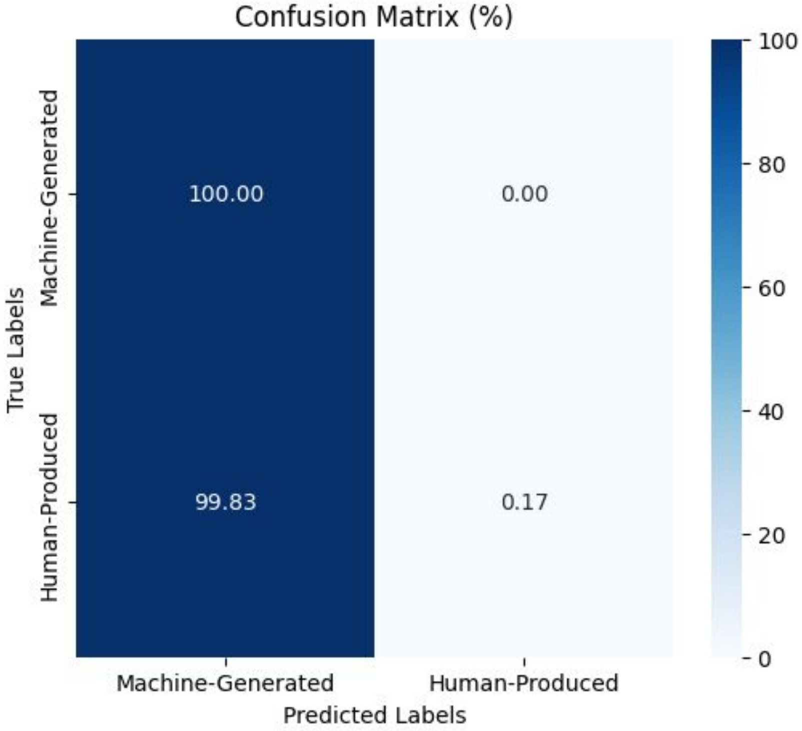
Accuracy: 0.5025
Precision: 0.5013
Recall: 1.0000

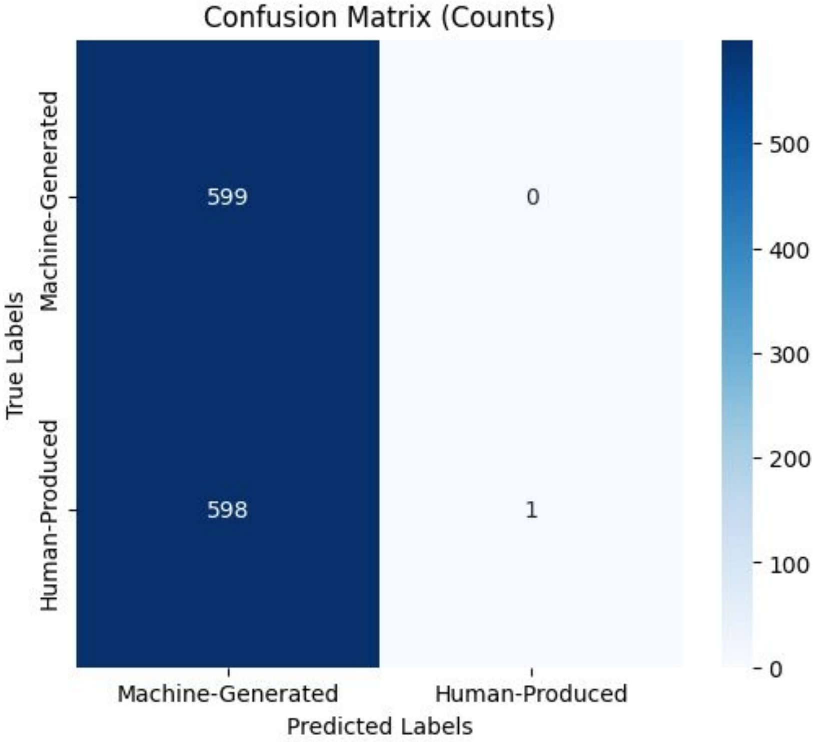




Testing with wikipedia_chatgpt_test_with_\n_without_unicode

Accuracy: 0.5008
Precision: 0.5004
Recall: 1.0000


Confusion Matrix (%)

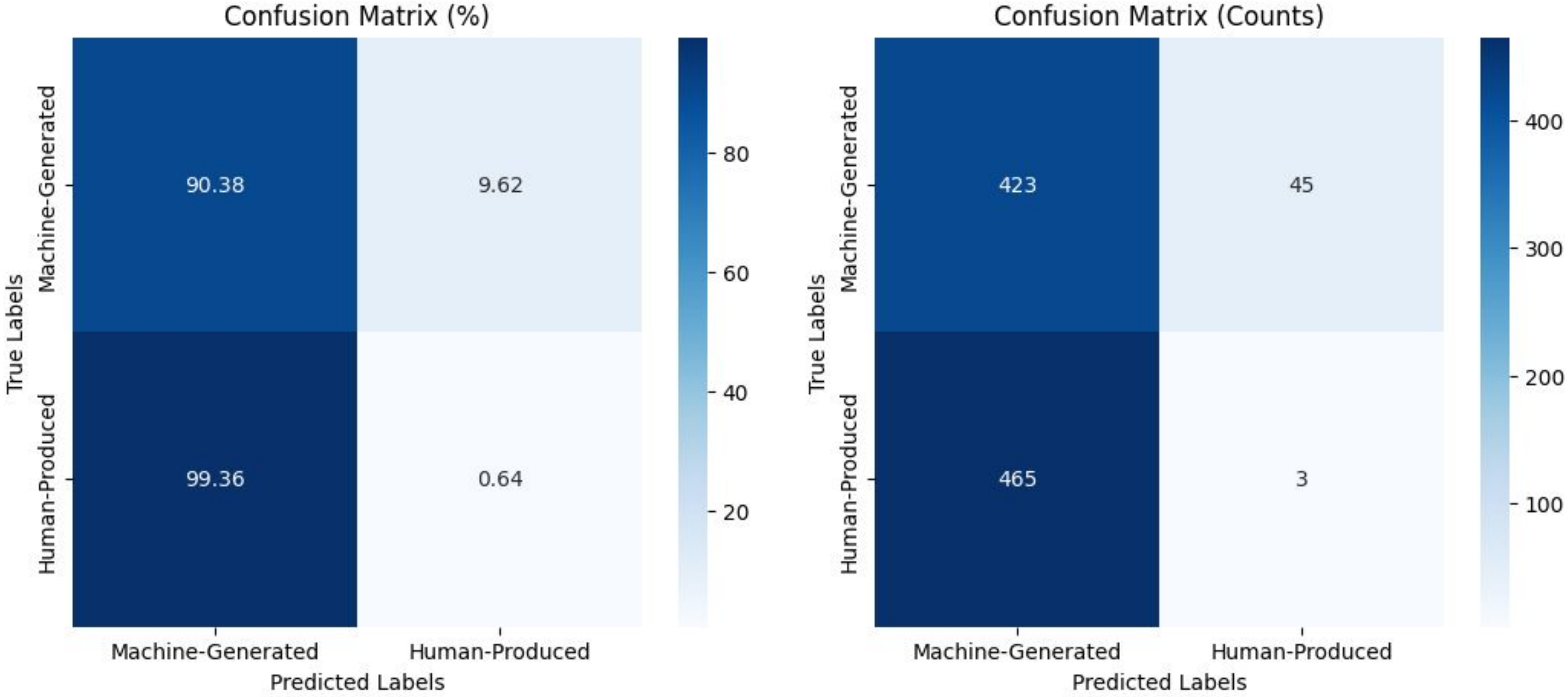

Confusion Matrix (Counts)

Testing with wikipedia_cohere_test_with_\n_without_unicode

Accuracy: 0.4551

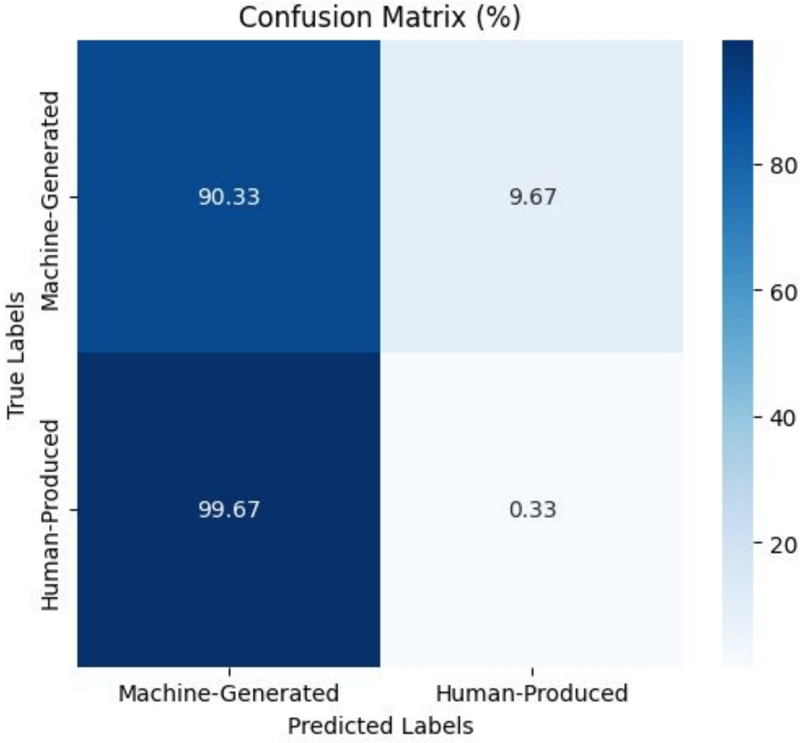
Precision: 0.4764

Recall: 0.9038

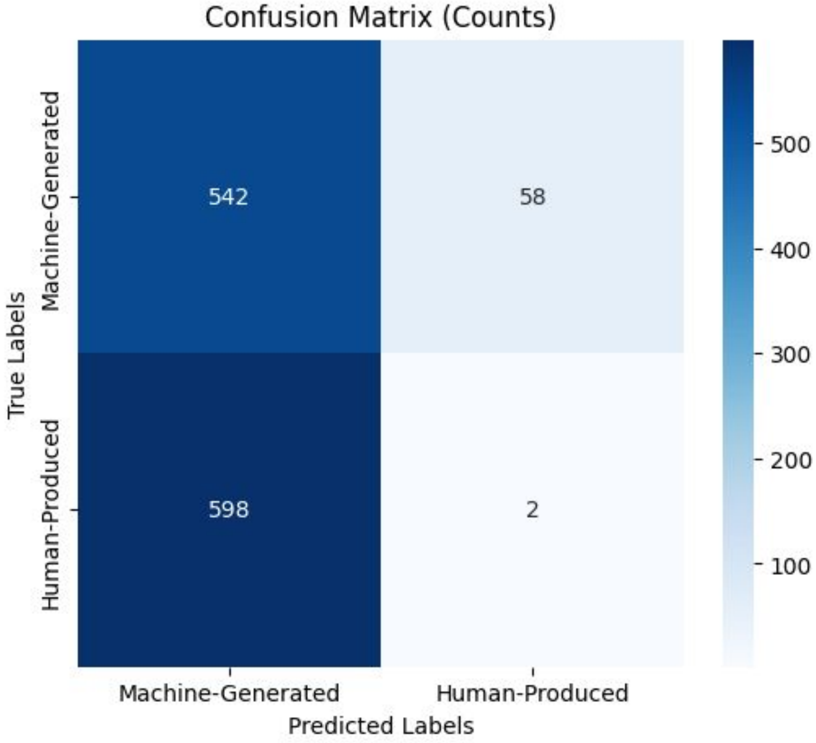


Testing with wikipedia_davinci_test_with_\n_without_unicode

Accuracy: 0.4533
Precision: 0.4754
Recall: 0.9033



Confusion Matrix (%)



Confusion Matrix (Counts)

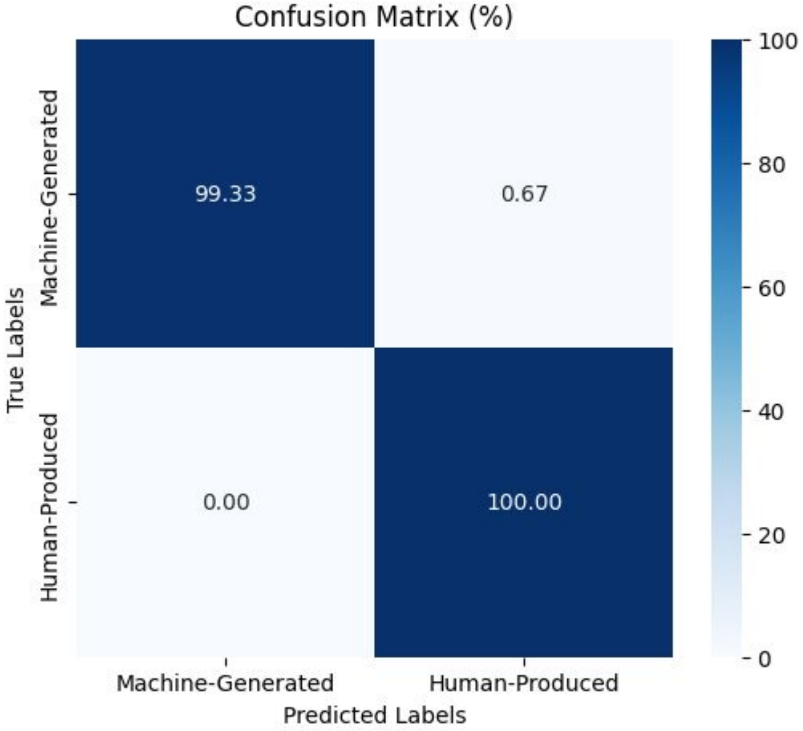
Bloomz-560m-academic-detector: **finetune using arxiv_bloomz_\n_removed + wikipedia_bloomz_with_\n**

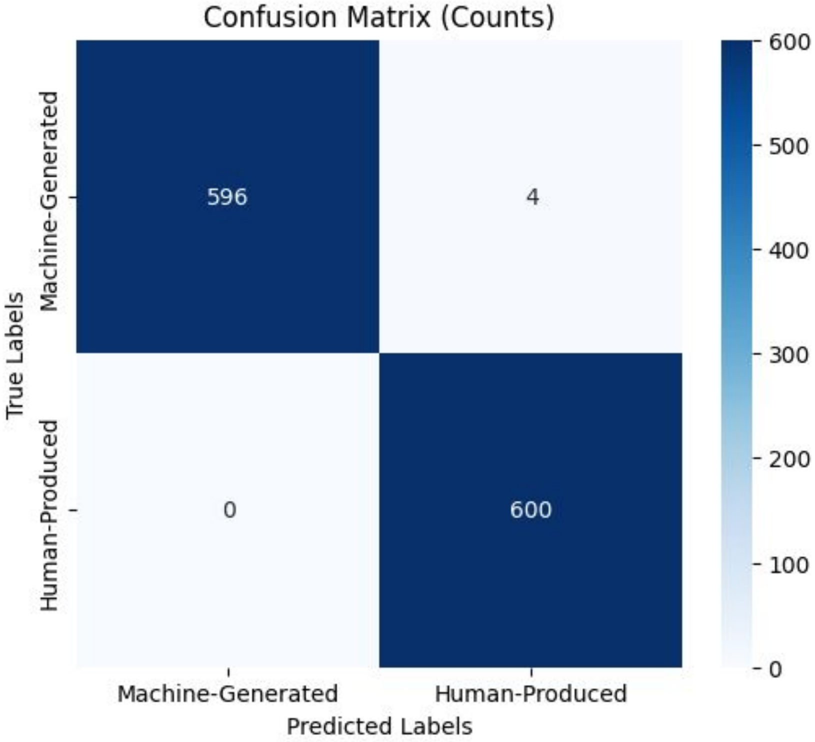
Test using:

arxiv_\n_removed_(chatGPT,cohere,flant5,davinci,bloomz)_test_set

Testing with wikipedia_bloomz_with_\n held-out test_dataset

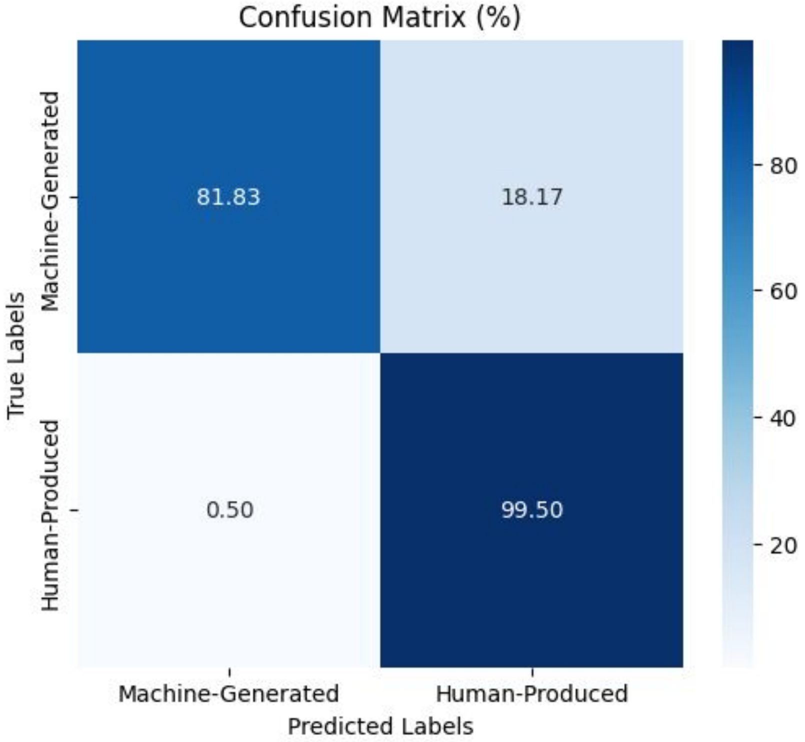
Accuracy: 0.9967
Precision: 1.0000
Recall: 0.9933

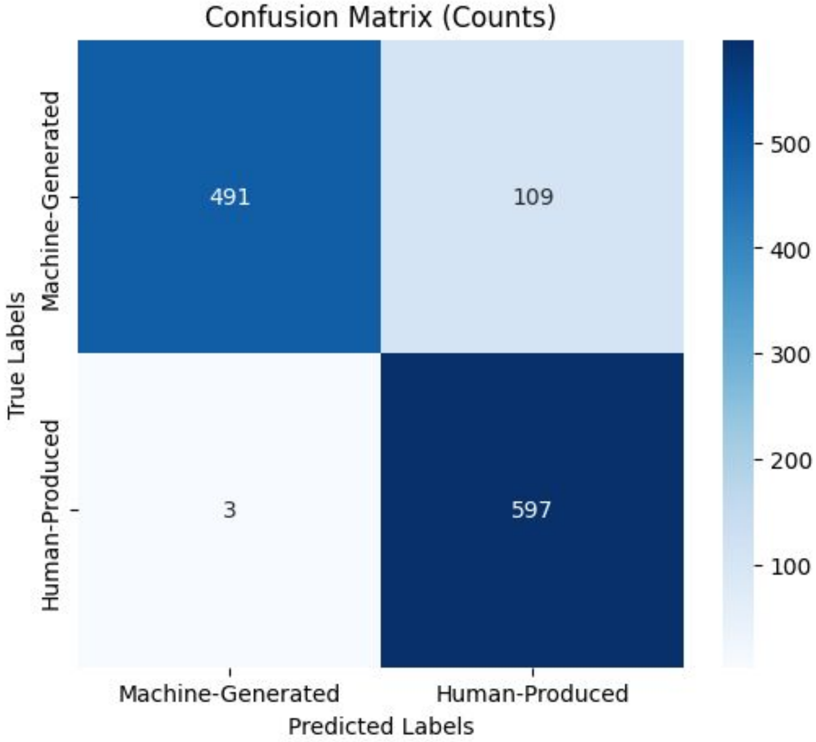




Testing with arxiv_bloomz_\n_removed_test

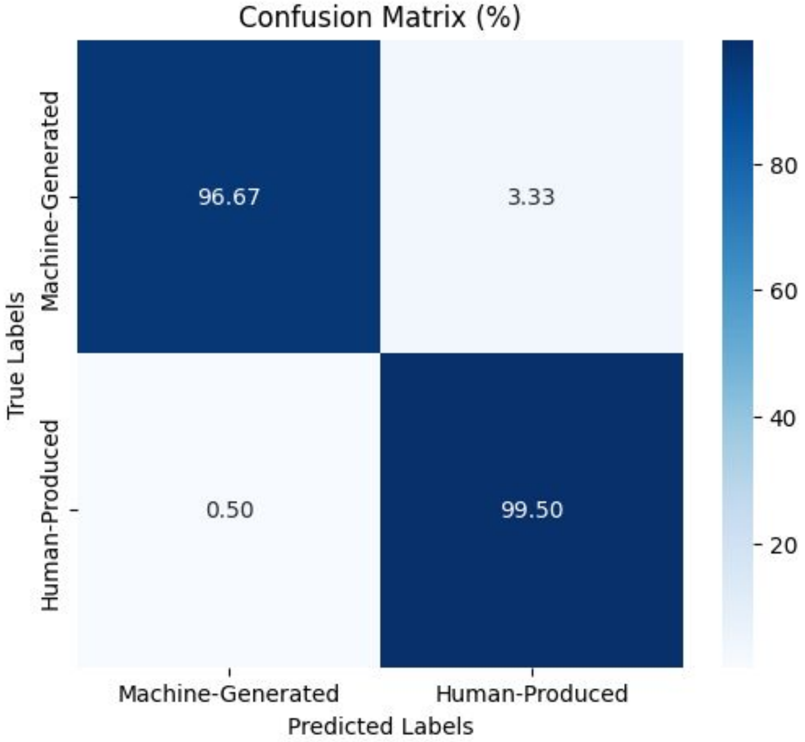
Accuracy: 0.9067
Precision: 0.9939
Recall: 0.8183

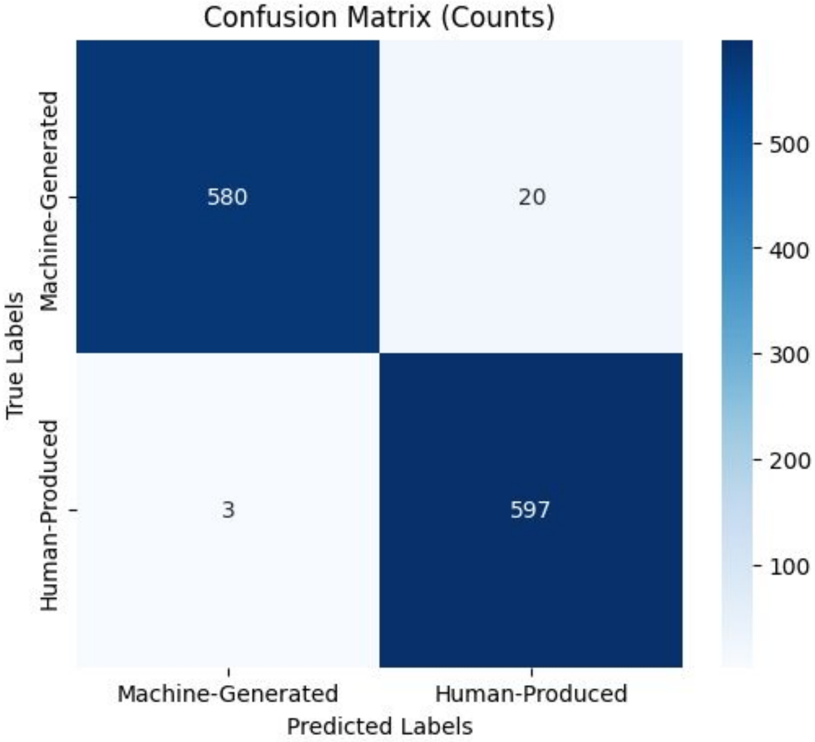




Testing with arxiv_ChatGPT_\n_removed_test

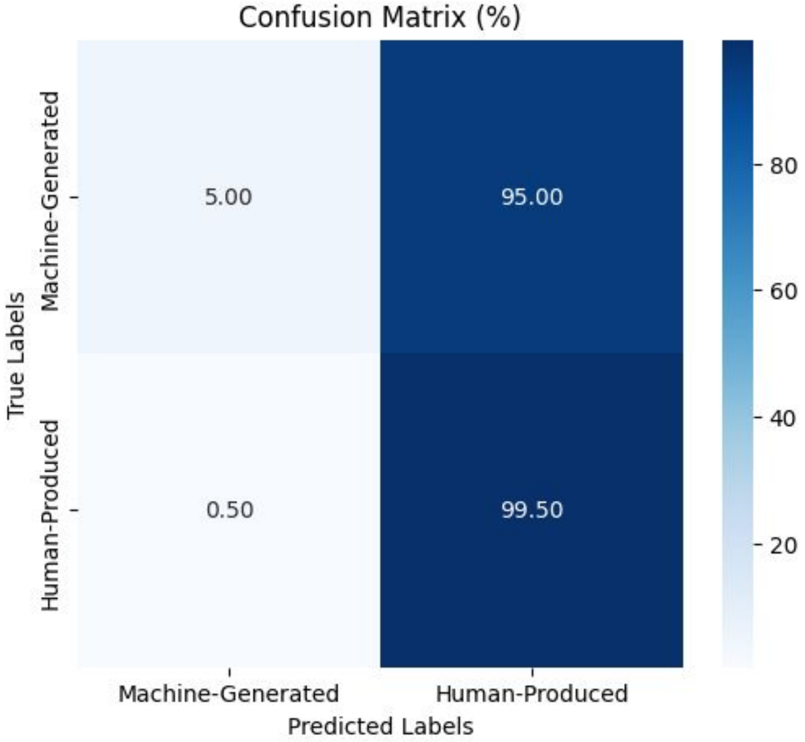
Accuracy: 0.9808
Precision: 0.9949
Recall: 0.9667

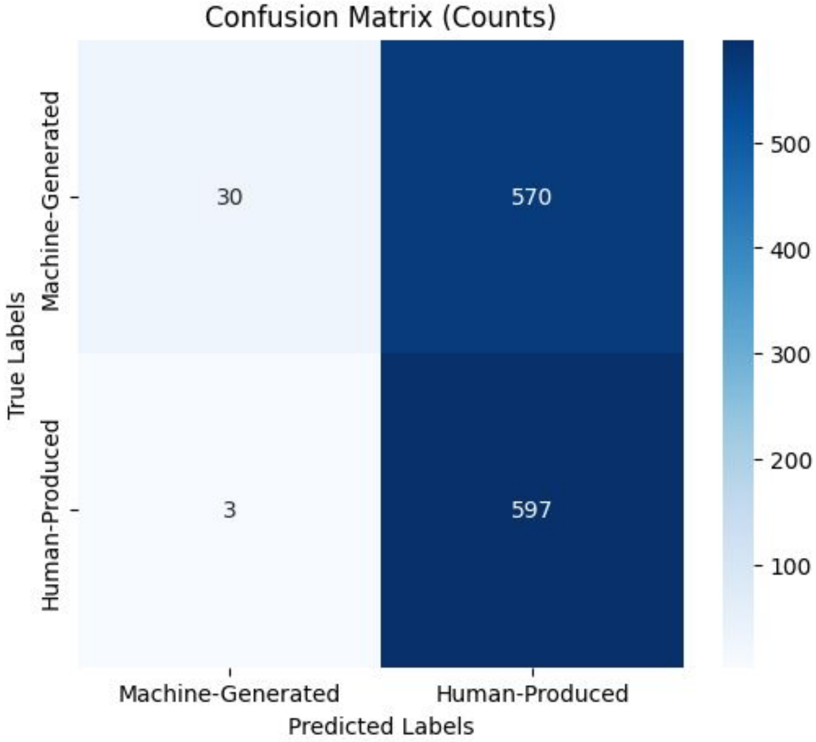




Testing with arxiv_cohere_\n_removed_test

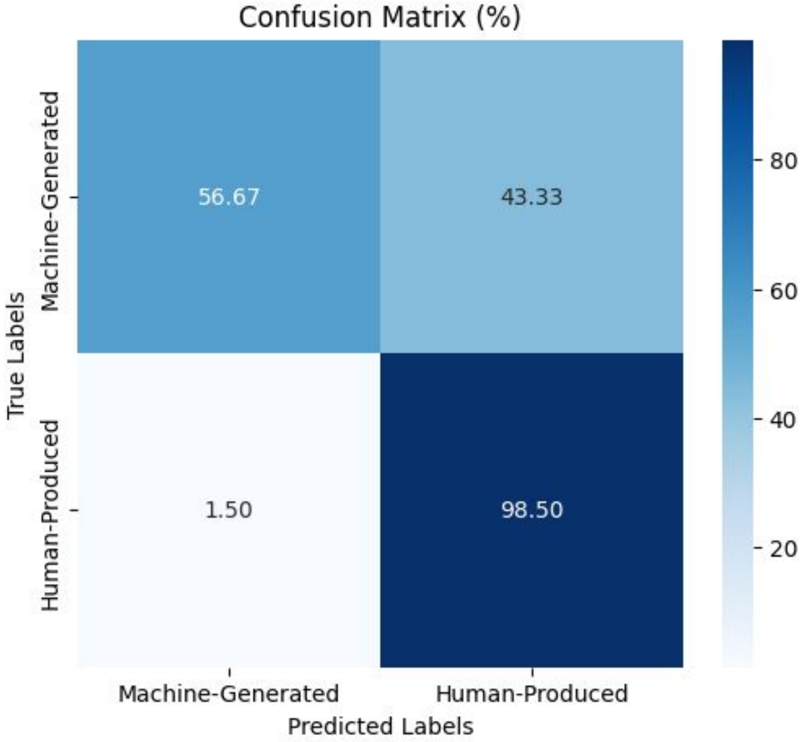
Accuracy: 0.5225
Precision: 0.9091
Recall: 0.0500

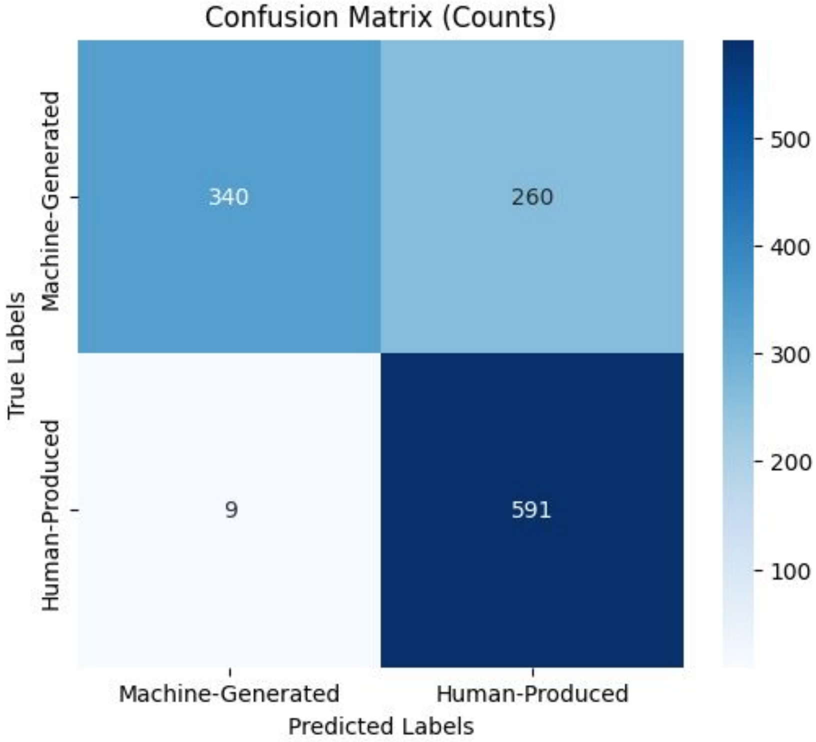




Testing with **arxiv_davinci_\n_removed_test**

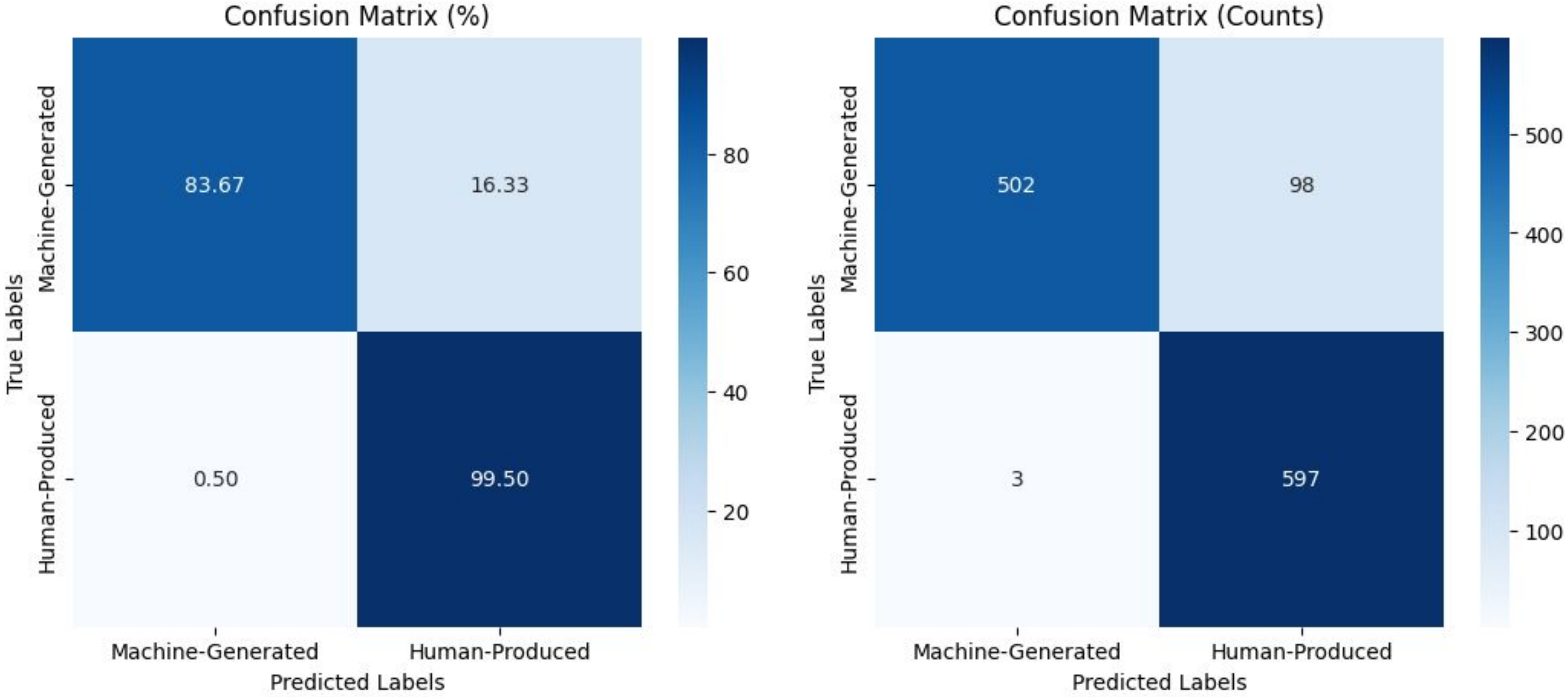
Accuracy: 0.7758
Precision: 0.9742
Recall: 0.5667





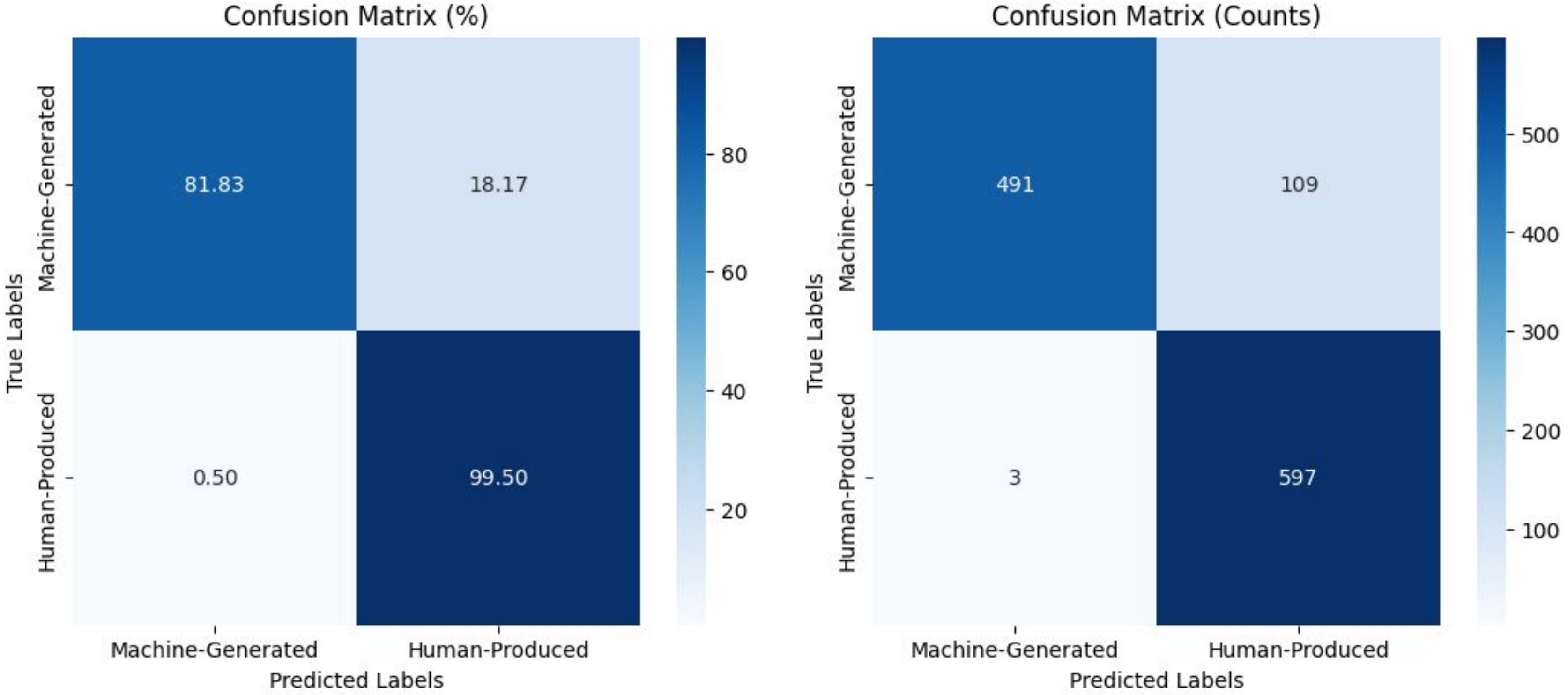
Testing with arxiv_flant5_\n_removed_test

Accuracy: 0.9158
Precision: 0.9941
Recall: 0.8367



Testing with arxiv_bloomz_paraphrased_\n_removed_test

Accuracy: 0.9067
Precision: 0.9939
Recall: 0.8183

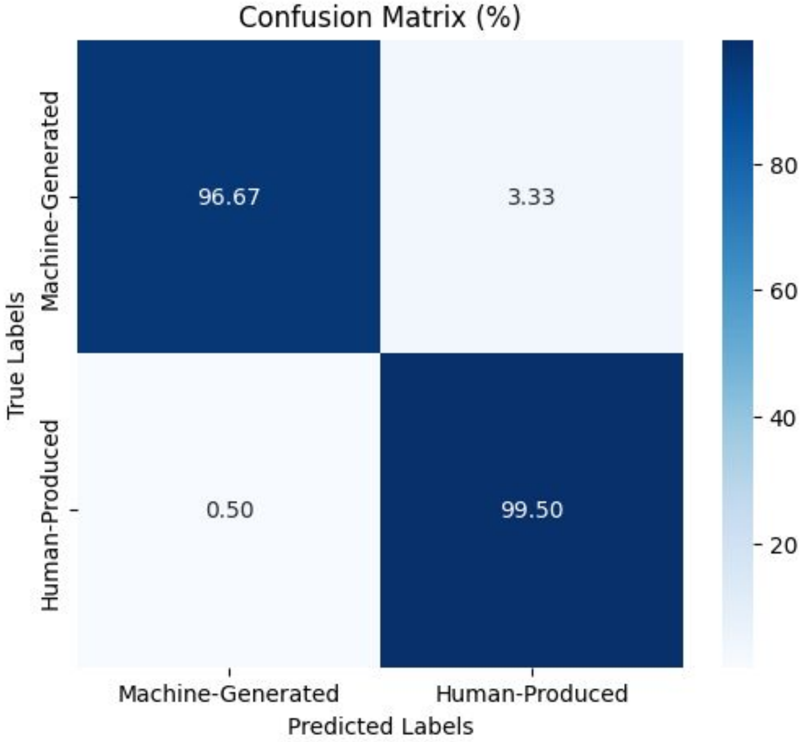


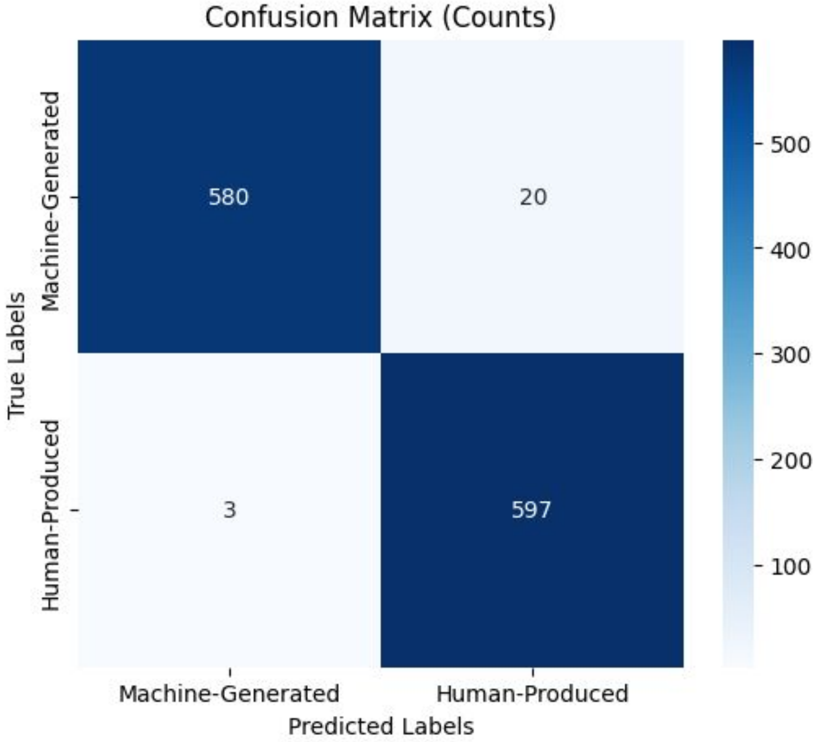
Testing with arxiv_chatGPT_paraphrased_\n_removed_test

Accuracy: 0.9808

Precision: 0.9949

Recall: 0.9667


Confusion Matrix (%)
True Labels
Machine-Generated
Human-Produced
Machine-Generated
Human-Produced
Predicted Labels


Confusion Matrix (Counts)
True Labels
Machine-Generated
Human-Produced
Machine-Generated
Human-Produced
Predicted Labels