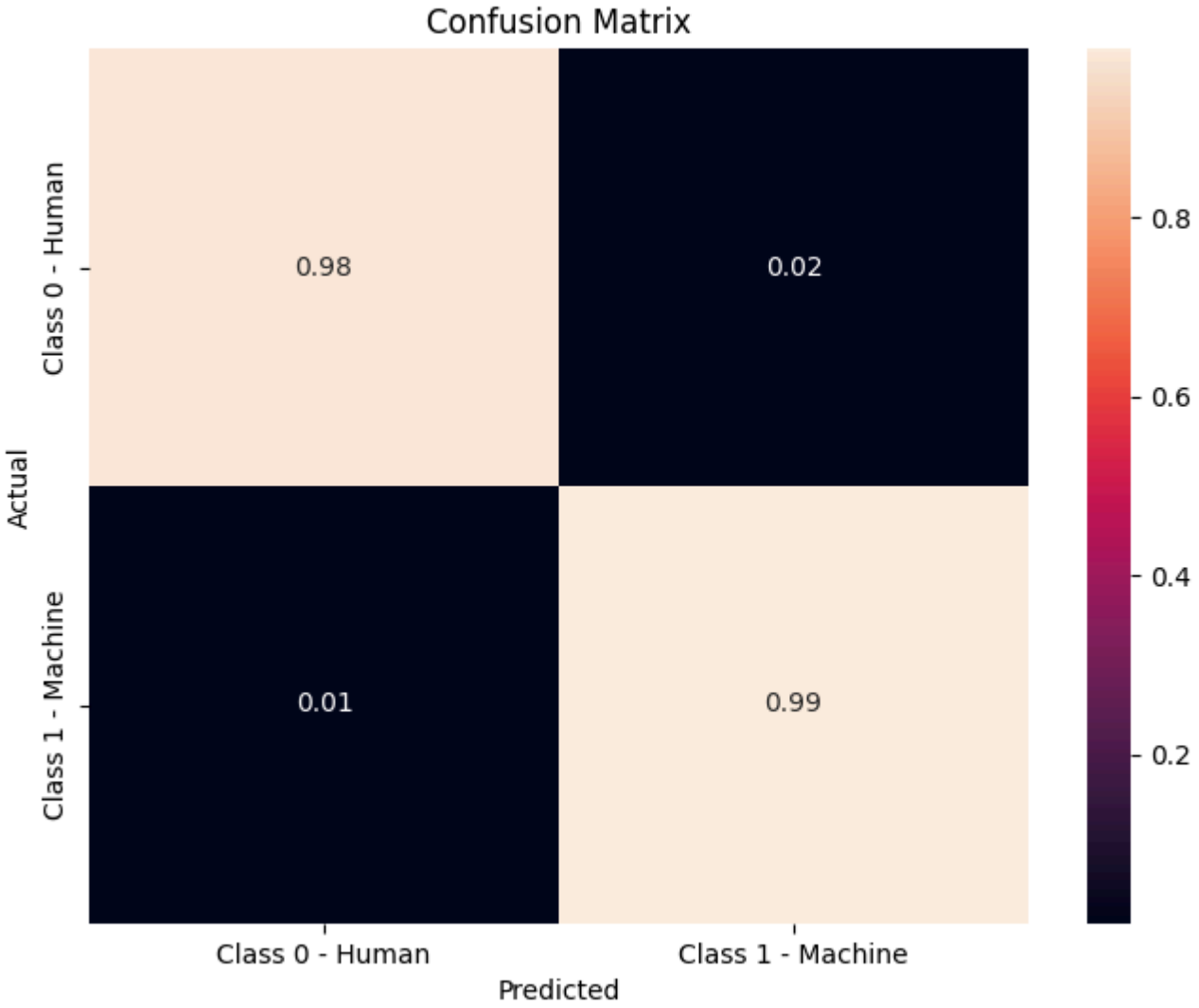
Report

Key	Value
Model Type	roberta
Task	evalPretrained
Training Data	chatgpt
Data Type	abstract
Preprocessing	False
Log Folder Name	roberta_evalPretrained_withn
Title	Roberta Pretrained Evaluation without preprocessing the generations \n
percentage	None
log_path	results/report/roberta/roberta_evalPretrained_withn//
Evaluation for chatgpt_abstract_with	{'test_loss': 0.00013447794481180608, 'test_model_preparation_time': 0.0033, 'test_accuracy': 1.0, 'test_precision': 1.0, 'test_recall': 1.0, 'test_f1': 1.0, 'test_sensitivity': 1.0, 'test_specificity': 1.0, 'test_runtime': 37.4559, 'test_samples_per_second': 32.038, 'test_steps_per_second': 2.002}
Evaluation for bloomz_abstract_with	{'test_loss': 2.581085681915283, 'test_model_preparation_time': 0.0037, 'test_accuracy': 0.6025, 'test_precision': 1.0, 'test_recall': 0.205, 'test_f1': 0.3402489626556016, 'test_sensitivity': 0.205, 'test_specificity': 1.0, 'test_runtime': 39.1322, 'test_samples_per_second': 30.665, 'test_steps_per_second': 1.917}
Evaluation for cohere_abstract_with	{'test_loss': 1.6555068492889404, 'test_model_preparation_time': 0.0035, 'test_accuracy': 0.7241666666666666, 'test_precision': 1.0, 'test_recall': 0.4483333333333333, 'test_f1': 0.619102416570771, 'test_sensitivity': 0.4483333333333333, 'test_specificity': 1.0, 'test_runtime': 39.8125, 'test_samples_per_second': 30.141, 'test_steps_per_second': 1.884}
Evaluation for davinci_abstract_with	{'test_loss': 1.7772784233093262, 'test_model_preparation_time': 0.008, 'test_accuracy': 0.7141666666666666, 'test_precision': 1.0, 'test_recall': 0.42833333333333334, 'test_f1': 0.5997666277712953, 'test_sensitivity':

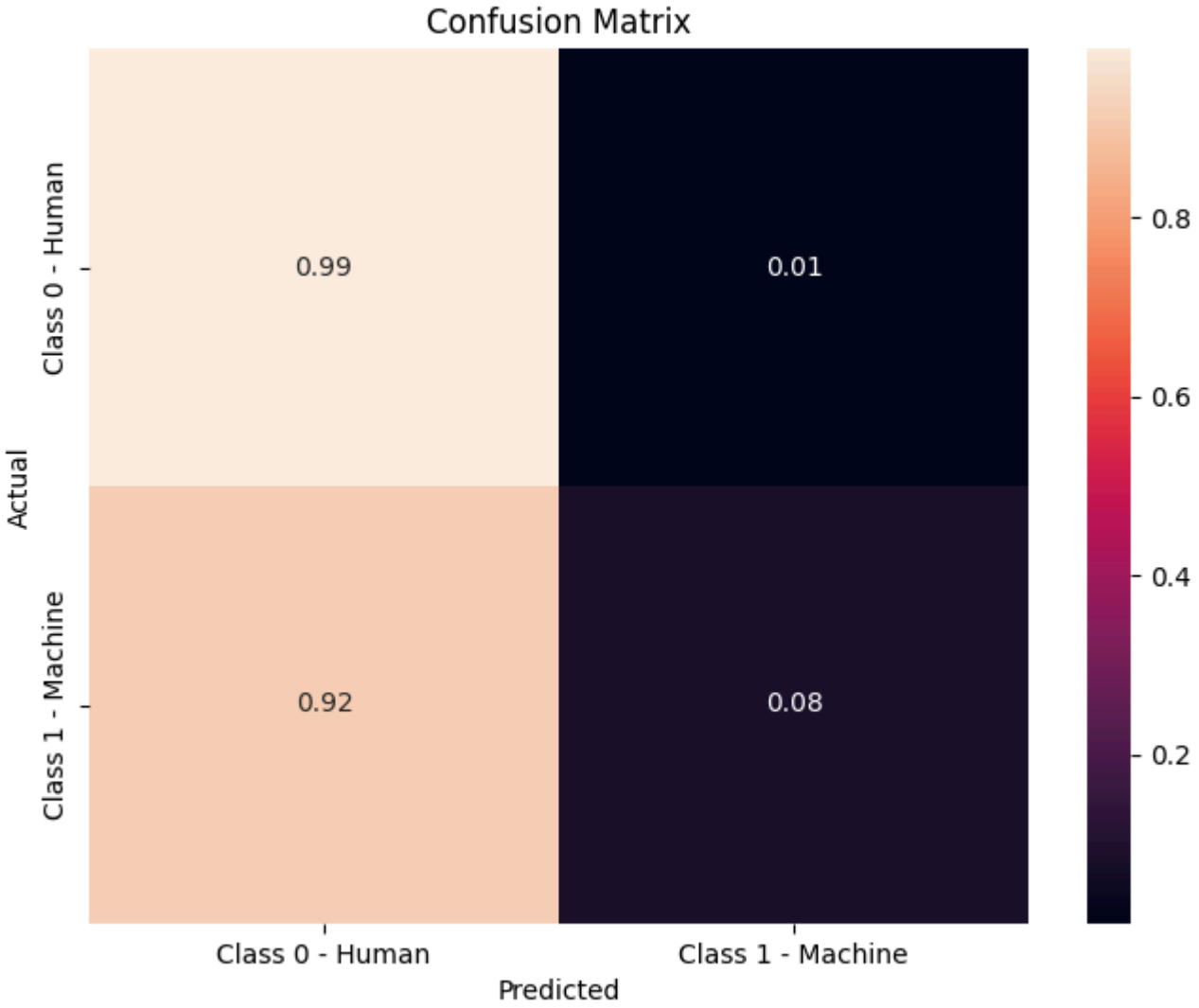
	0.42833333333333334, 'test_specificity': 1.0, 'test_runtime': 38.868, 'test_samples_per_second': 30.874, 'test_steps_per_second': 1.93}
Evaluation for flant5_abstract_with	{'test_loss': 0.8194342255592346, 'test_model_preparation_time': 0.0051, 'test_accuracy': 0.8641666666666666, 'test_precision': 1.0, 'test_recall': 0.7283333333333334, 'test_f1': 0.8428158148505304, 'test_sensitivity': 0.7283333333333334, 'test_specificity': 1.0, 'test_runtime': 38.88, 'test_samples_per_second': 30.864, 'test_steps_per_second': 1.929}
Evaluation for llama3_ml_with	{'test_loss': 0.13231664896011353, 'test_model_preparation_time': 0.0034, 'test_accuracy': 0.9733333333333334, 'test_precision': 0.9701986754966887, 'test_recall': 0.9766666666666667, 'test_f1': 0.9734219269102989, 'test_sensitivity': 0.9766666666666667, 'test_specificity': 0.97, 'test_runtime': 38.8083, 'test_samples_per_second': 30.921, 'test_steps_per_second': 1.933}
Evaluation for bloomz_wiki_with	{'test_loss': 3.0643136501312256, 'test_model_preparation_time': 0.0057, 'test_accuracy': 0.535, 'test_precision': 0.875, 'test_recall': 0.08166666666666667, 'test_f1': 0.14939024390243902, 'test_sensitivity': 0.08166666666666667, 'test_specificity': 0.9883333333333333, 'test_runtime': 39.0392, 'test_samples_per_second': 30.738, 'test_steps_per_second': 1.921}
Evaluation for chatgpt_wiki_with	{'test_loss': 0.0754389613866806, 'test_model_preparation_time': 0.0033, 'test_accuracy': 0.9858096828046744, 'test_precision': 0.9833887043189369, 'test_recall': 0.988313856427379, 'test_f1': 0.9858451290591174, 'test_sensitivity': 0.988313856427379, 'test_specificity': 0.9833055091819699, 'test_runtime': 41.1151, 'test_samples_per_second': 29.138, 'test_steps_per_second': 1.824}
Evaluation for cohere_wiki_with	{'test_loss': 1.4417712688446045, 'test_model_preparation_time': 0.0032, 'test_accuracy': 0.7329059829059829, 'test_precision': 0.9780701754385965, 'test_recall': 0.47649572649572647, 'test_f1': 0.6408045977011495, 'test_sensitivity': 0.47649572649572647, 'test_specificity': 0.9893162393162394, 'test_runtime': 32.1792, 'test_samples_per_second': 29.087, 'test_steps_per_second': 1.833}
Evaluation for davinci_wiki_with	{'test_loss': 0.5457635521888733, 'test_model_preparation_time': 0.0031, 'test_accuracy': 0.8983333333333333, 'test_precision': 0.9838056680161943, 'test_recall': 0.81, 'test_f1': 0.8884826325411336, 'test_sensitivity': 0.81, 'test_specificity': 0.9866666666666667, 'test_runtime': 41.856, 'test_samples_per_second': 28.67, 'test_steps_per_second': 1.792}

Plots/Confusion Matrix

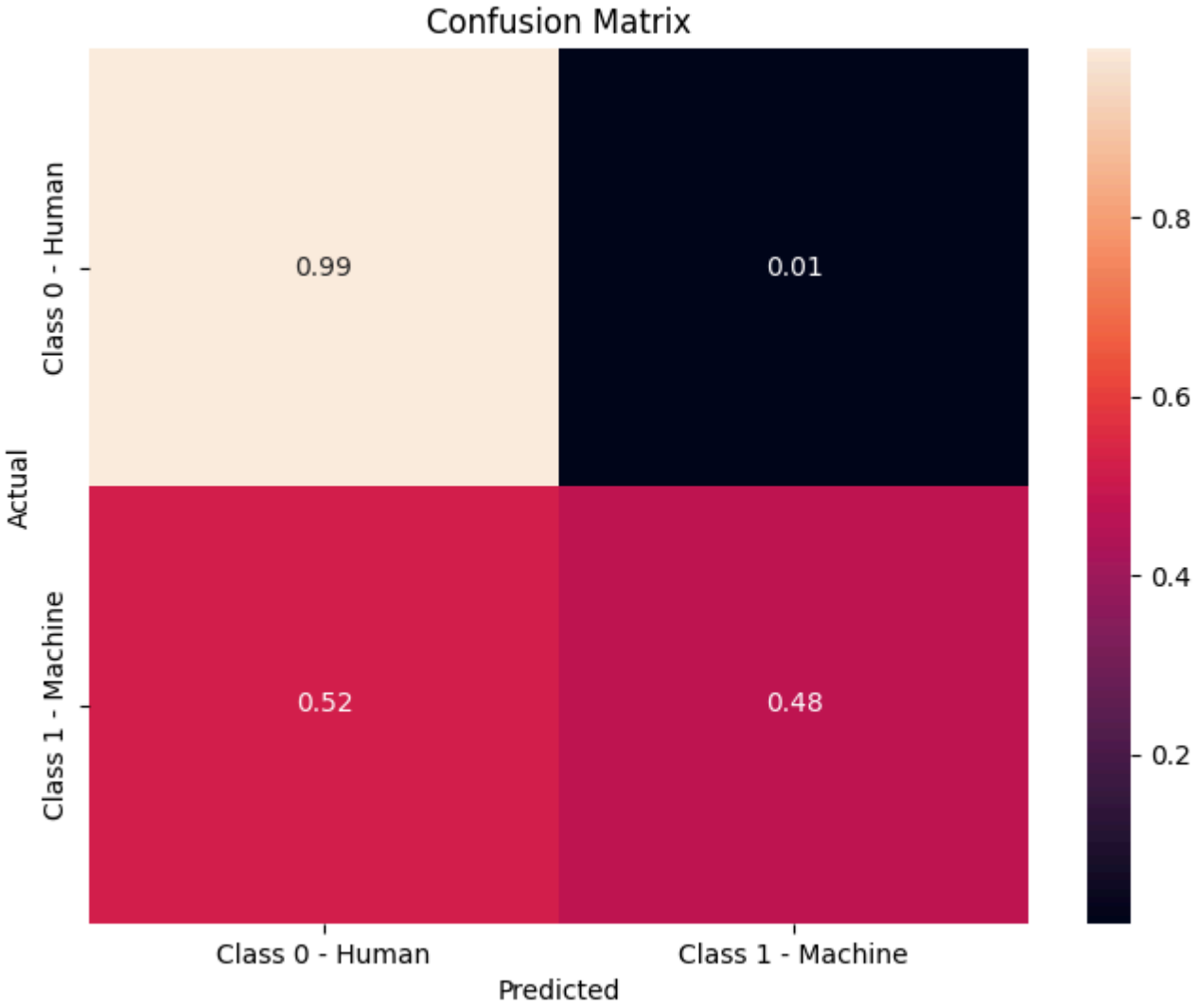
Evaluation on chatgpt wiki generations



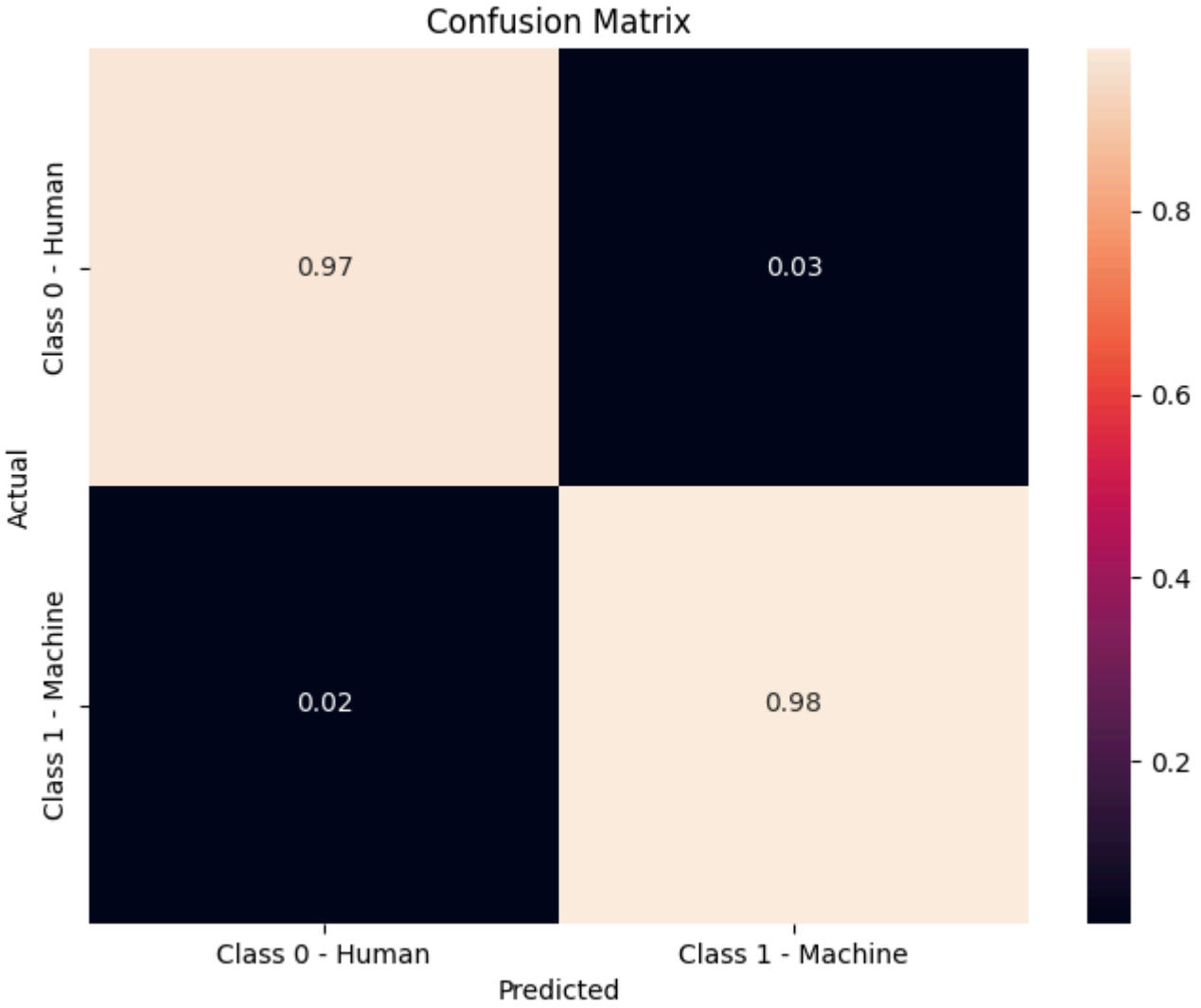
Evaluation on bloomz wiki generations



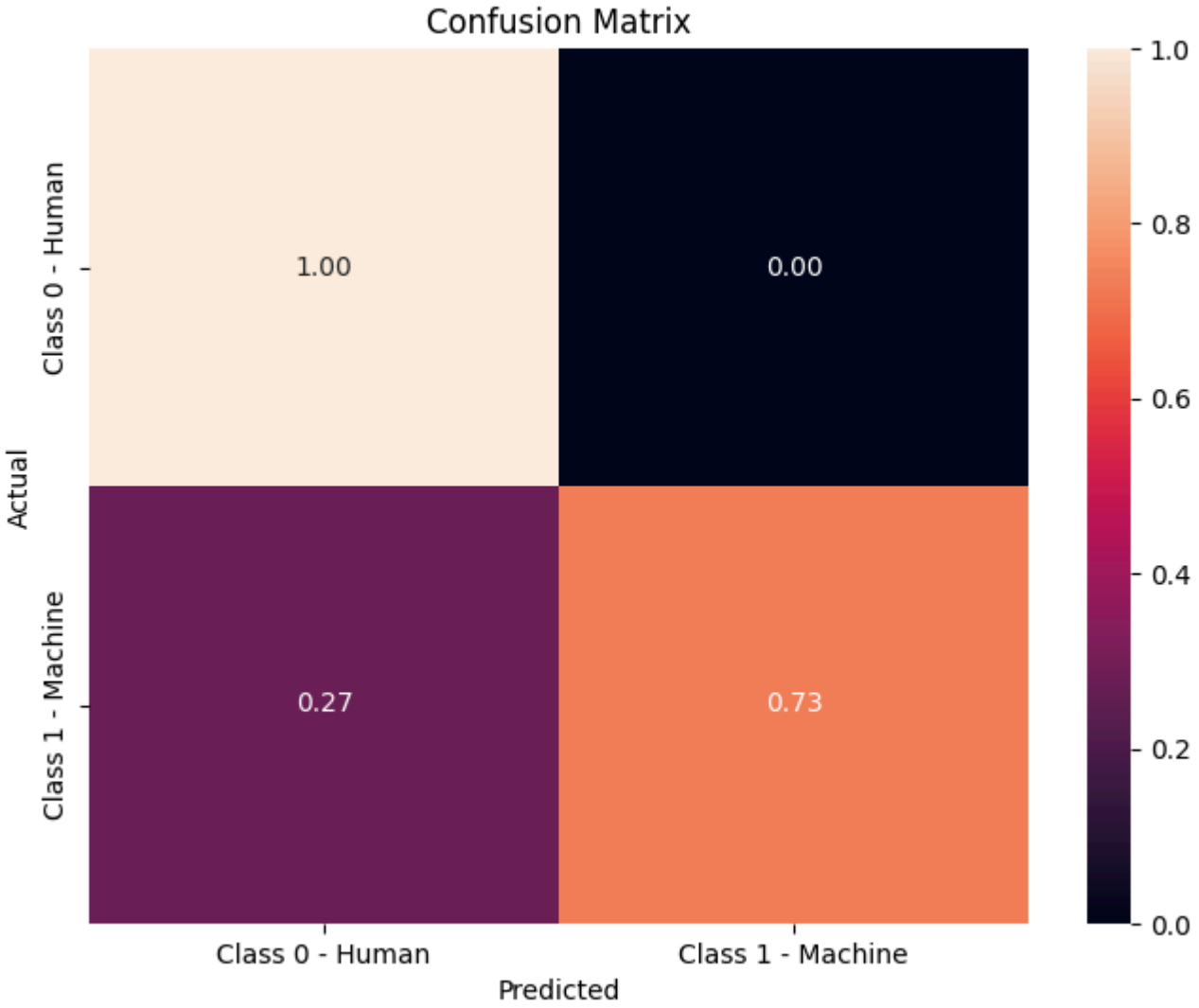
Evaluation on cohere wiki generations



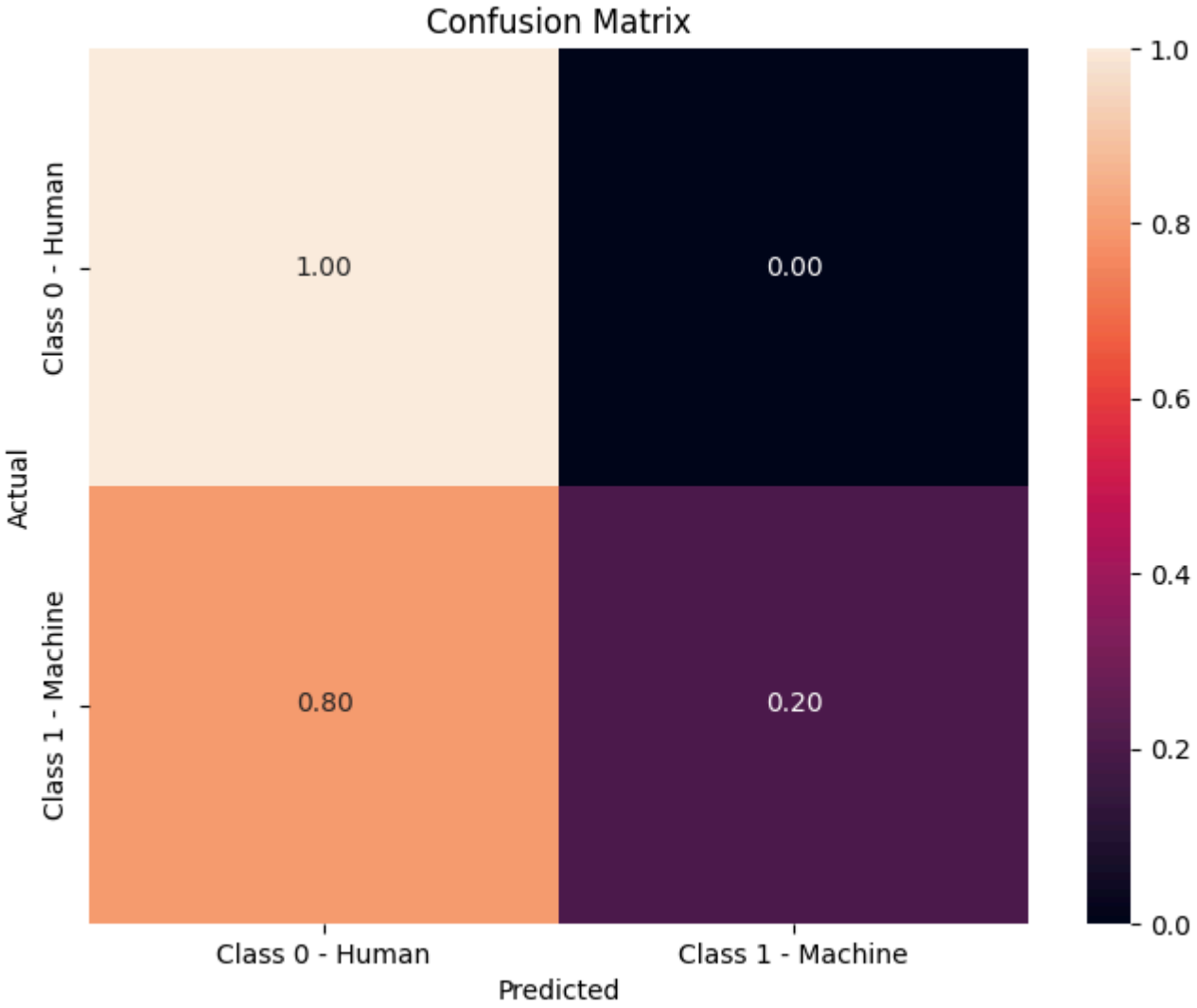
Evaluation on llama3 wiki generations



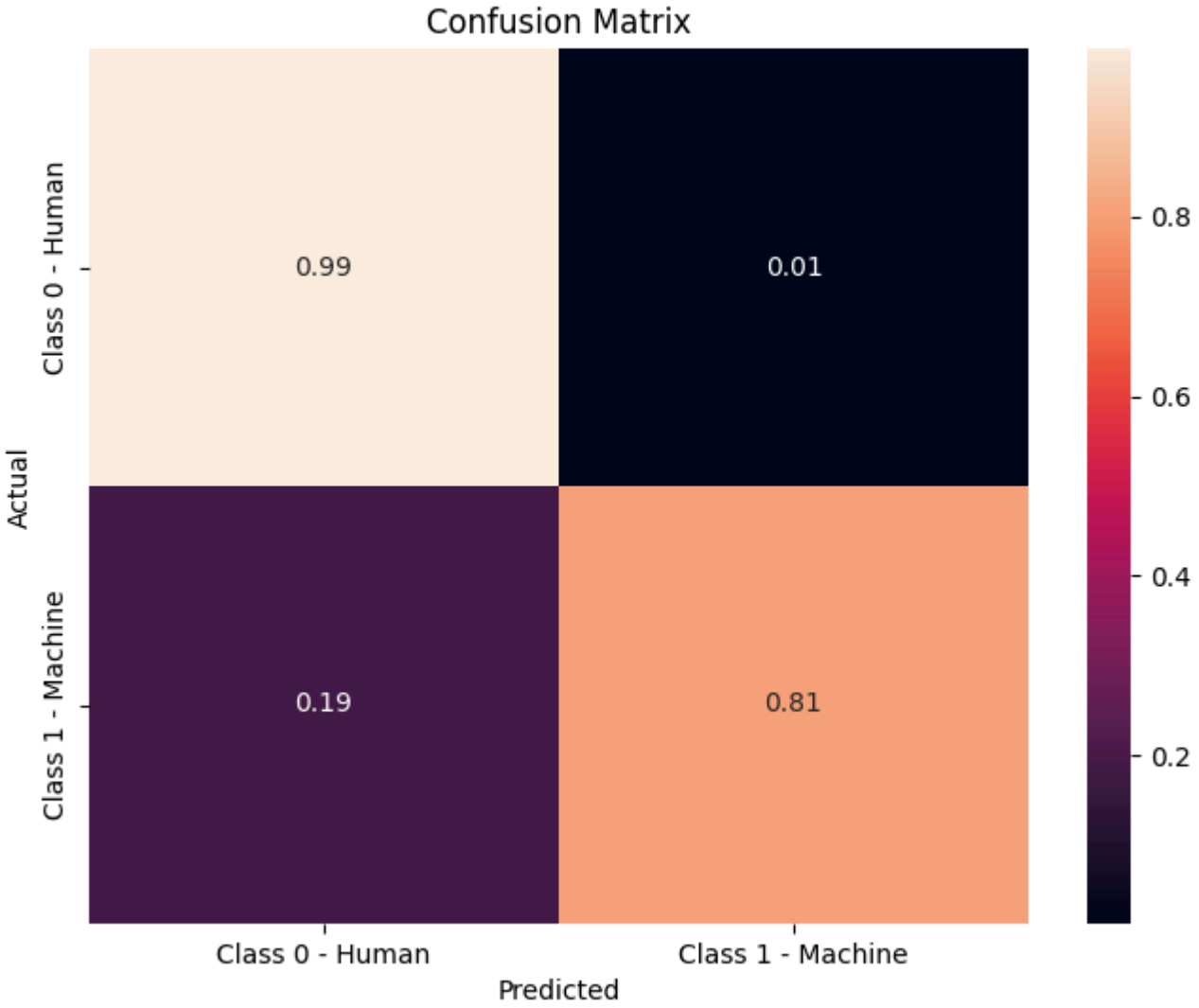
Evaluation on flant5 abstracts generations



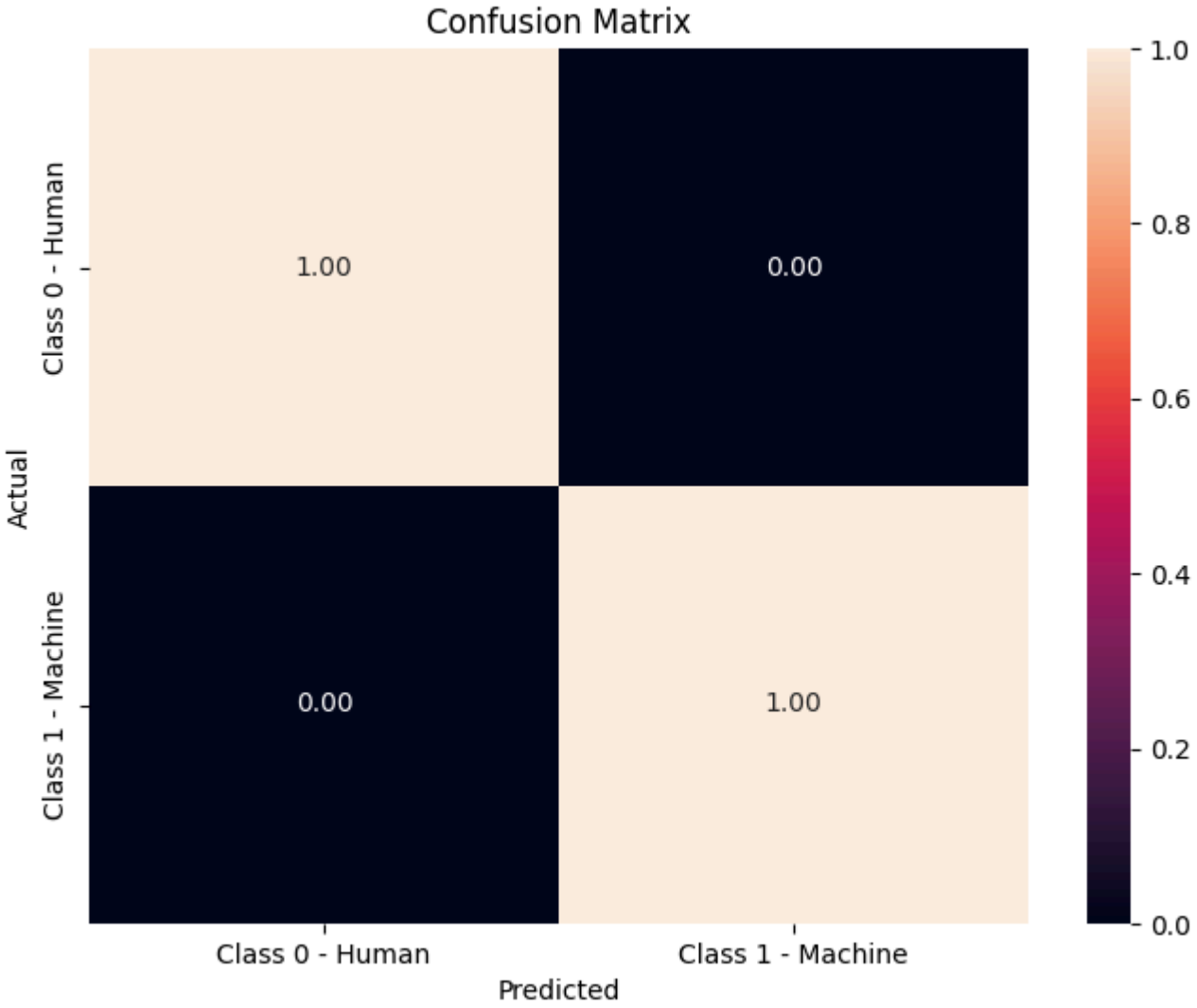
Evaluation on bloomz abstract generations



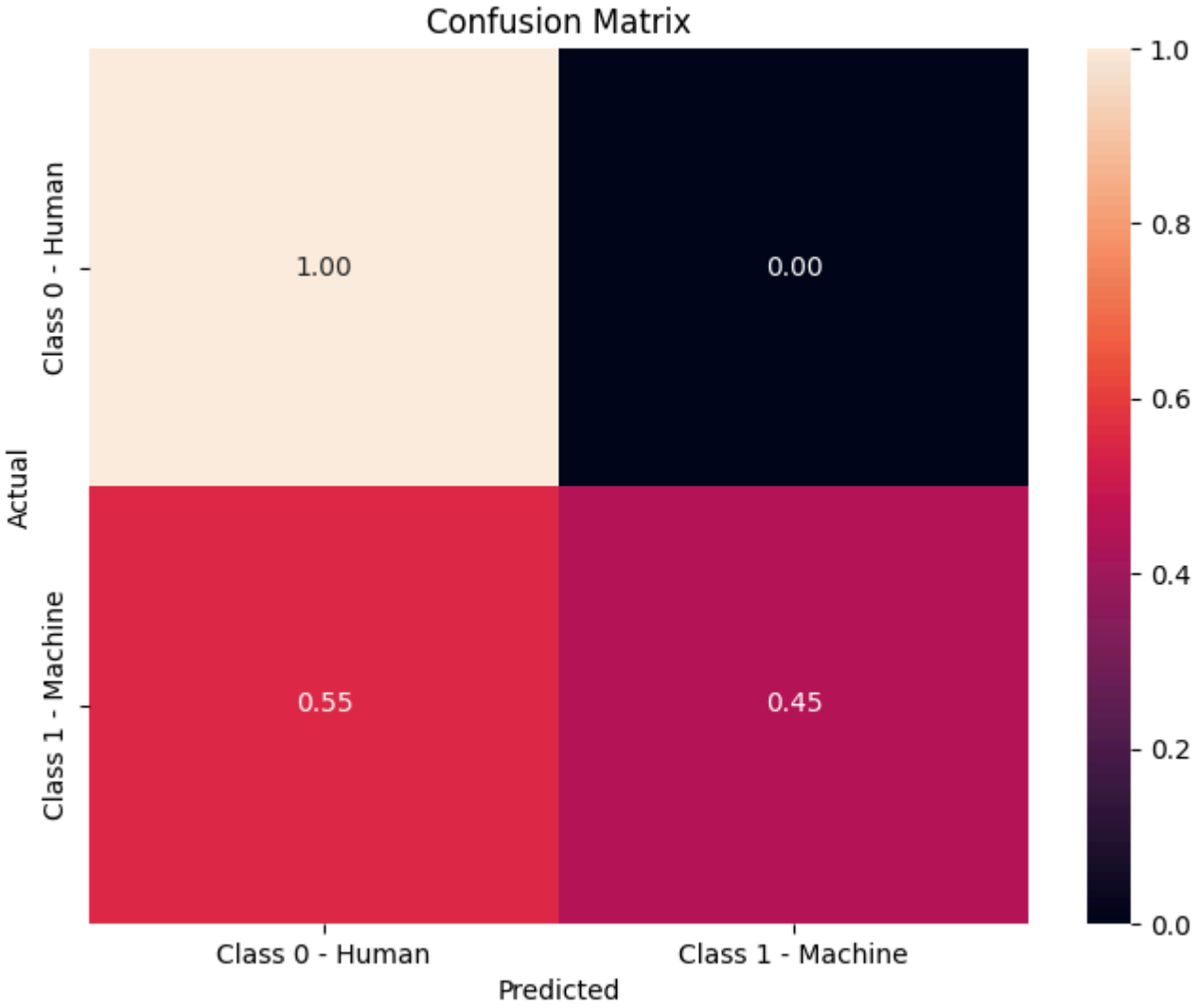
Evaluation on davinci wiki generations



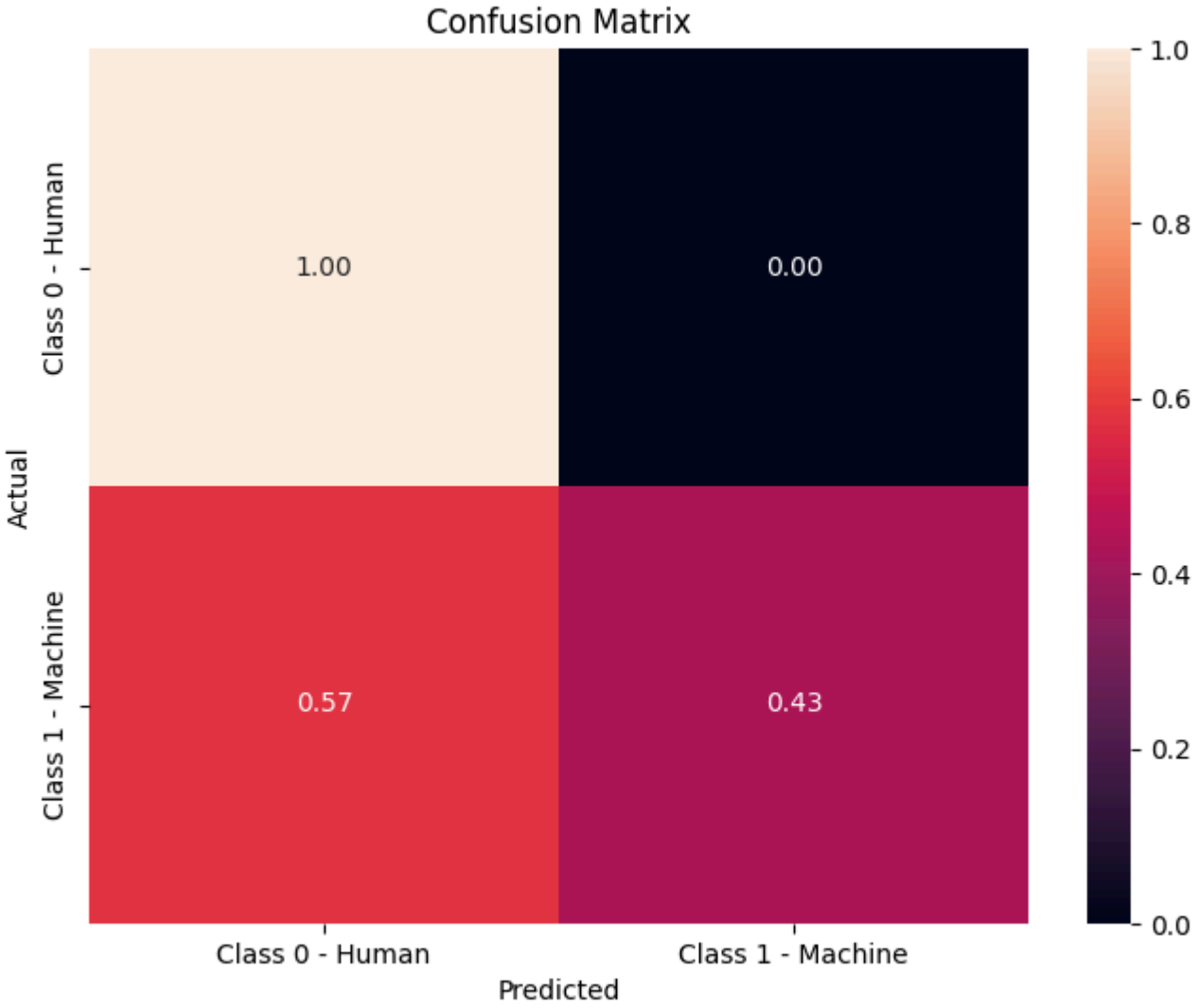
Evaluation on chatgpt abstract generations



Evaluation on cohere abstract generations



Evaluation on davinci abstract generations



Original text Vs Perturbed Text

Attack Summary