Detect AI Generated Text Dataset

### Synthetic Essay Generation

We used an internal tool for synthetic data generation followed by a custom script for data cleaning and statistical matching.

#### Models

Palm 2 for Text (text-bison)\*

Palm 2 for Text (text-unicorn)\*

Gemini 1.0 Pro\*

Gemini 1.0 Ultra\*

\* Public availability on [Vertex AI](https://cloud.google.com/vertex-ai/generative-ai/docs/learn/models)

Non-instruction-tuned variants of each of the above, not publicly available.

GPT 4

#### Sampling

Random sampling with temperature 1.0-1.6. One example parsed per model call.

#### Model Input

Three-shot prompting with examples randomly selected from student essays of the same topic. Examples were selected randomly before each model call (each generated essay got its own three random student essays). See Appendix 1 for an example.

#### Prompt Template

text: {example 1}

text: {example 2}

text: {example 3}

Read the following article: """{article}"""

{prompt}

text: {

### Post-Processing

1. Duplicate removal: [Universal Sentence Encoder](https://www.tensorflow.org/hub/tutorials/semantic_similarity_with_tf_hub_universal_encoder) embedding cosine similarity at 85% threshold.
2. Manual review for pathological outputs. We removed any generated texts that obviously would not be submitted by students, e.g., computer code.
3. Cleaned real and synthetic essays using a common script: removed spurious formatting, special characters, non-standard punctuation, placeholders. See Appendix 2.
4. Matched text distribution of generated essays to student essays across Word Count, Sentence Length, Function Word Count. See Appendix 2.

In all, we retained maybe 10% of the original number of generated essays.

## Appendix 1 - Example Prompt

text: {Dear, State senator Voting is a public thing that happens every fourth year when a President is resining or cannidating for a second run. People all around the United States vote for a President they think can run the country the best. I think it's not fair that our votes are getting thrown away, and not even looked at. bradford Plumer said that "the Electoral College is unfair, outdated, and irrational source 2 paragraph 14, and i think that is true. I think we should Bannish the Electoral College because it is a old process, unfair to Americans, and can lead to disasters.

The Electoral College is a very outdated process. Dont you think we should use a new one? The electoral College has been use since the 13 colonies, and has been ignoring the public ever since. I think its only fair if we have a part in the president election. Also, if we could change it to popular vote, more people would be intact with politics. Bradford Plumer says that "the election is only a few swing voters away from a catastrophe. source 2 paragraph 12. I agree with him because the Electoral College is very old and can break at any time. I think its time that we change that and start fresh.

America's saying is "home of the free, land of the brave" well its not free if we cant vote for our president we want, and the government does it for us. Bradford Plumer says that the electoral college is unfair to voters, and candidates because of the winner takes all system source 2 paragraph 13. The winner takes all system is states choosing if the are going to pick a republican president or demicrate president. Lastly. it is unfair to us. We have no power against the government, and our votes mean nothing to them.

The Electoral College can also lead to disasters! rivets can flood the streets if they dont like the legislators choice, and that can lead to deaths. Secondly, shootings could start to occurring around the states, and can get out of hand. Inicent silvilians can get hurt for no reasons. Lastly, there can be financial crashes, and everyone around the states could go in dept.

to sum up my letter i think the electoral College should be bannished for good. You have heard that it is a old process and can lead to falure, it is unfair to americans, and it can lead to disasters all around the country.}

text: {"Boys and girls of every age, wouldn't you like to see something strange?" After hundreds of years with the Electoral College working just fine, it would be very strange for Americans to just give it up in the name of convenience. The Electoral College should stay in place because our founding fathers created it, it keeps socialism and communism out, and it is in the constitution.

Our founding fathers did many things that would ensure the United States is in safe position for centuries to come. We have the freedom of speech, religion, press, to bear arms, and of privacy. These are just as important as the Electoral College. The founding fathers of the United States wanted to make sure Americans didn't vote for communist or socialist presidents by popular vote. So they created the Electoral College, a system of voting in which each state has a certain number of points based on the amount of Senators and Congressmen representing tha state. If a presidential candidate has a majority of votes in the state, the candidate wins all the points from the state. According to the first source, there is a current total of 538 points in America and it takes 270 for a candidate to win. There is not one thing the founding fathers did that harmed the United States America should have faith in these men and not charismatic candidates who plan to go against the constitution.

Furthermore, the Electoral college keeps socialism and communism out of our capitalist democracy. Sometimes people don't fully understand politics and vote to simply make their voice count somehow. It may sound illogical, but everyone makes wrong decisions sometimes. The Electoral College is in place to keep people from falling into traps and voting for socialist candidates who plan to rid the U.S. of its freedom. According to the second source, Jimmy Carter, one of the worst presidents America has ever had, and Bob Dole, a presidential candidate who lost to Bill Clinton, were in favor of getting rid of the Electoral College. If popular vote were in place, the one with the most votes overall wins. That would make it a popularity contest and anyone who has witnessed one of those in high school knows they never end well. America has not had a socialist president with the Electoral college who is to say there will not be one without it?

Additionally, the Electoral College is written in the constitution. According to the first source, the Electoral College in the constitution and is briefly mentioned in the 23rd amendment, which allows District of Colombia to participate in the Electoral College and gives it 3 electors. Getting rid of a law, or in this case a system of voting, from the constitution is never a logical decision. For example, the 8th amendment,"no cruel and unusual punishment", is not a wise choice to get rid of and nor is the 2nd amendment "the right to bear arms". Why? Because it would turn America into something it isn't it provides more government control over its citizens. Without the Electoral College, everyone's vote is seen and counted, even the independent ones and the votes of the elderly. The constitution is a guide for the people from those who fought and died for the country and it should never be altered for convinience.

In conclusion, the Electoral College should stay where it is because it was created by the founding fathers, it keeps communism and socialism out, and it is in the constitution. If the United States were to rely on anyone, it should be the founders of the country itself, for they were the ones who fought for the freedom of the country.}

text: {The electoral college is not a good thing and it certainly messes things up it takes votes away from people, It changes the votes to make a person lose presidency even by changing the senators, and it's a winner take all system these dont turn out good no matter what. It's even overthrowing the people with the popular vote yes it sadly is possible.

Why would you still have the electoral college if it doesnt help anyone at all?

There are many many ways to win an election and this is not one good way it can easily make one side lose even if they won the popular vote. The odds of one side winning is easily changed with electoral votes. This interrupts the odds of how much the people likes them then it turns into how much the government of states likes them. However it could help some people win if they are not the popular vote.

It can also make anyone with the popular vote lose when they should have won like after Al Gore during the gallop poll thanks to the perks of the electoral college won the popular vote but he lost presidency over it. Even though 60% of people prefer the direct election now some prefer the electoral college.

But this can all be due to the fact that the electoral system was made by our founding fathers on the Constitution so it may only be natural that they support it. A good argument was when the system allowed for far worse when the electors for defy the will of people like in passage 2 it says "in 1960 when segragationists in the louisiana legislature almost succeeded in replacing electors who would oppose John F. kennedy" that is just evil. Although the law has not been changed since the Constitution was made.

Even at the most basic level the electoral college is still unfair to voters because of the winner take all system put in place for each state. They saddly dont visit every state in passage 2 it says "Candidates dont spend time in states they know they have 0 chance of winning, focusing only on tight races" Although the winner take all is made just to do what they are using it for by awarding electors to the presidential candidate exept for Maine and Nebraska it is being used for what they are using it for.

In all the electoral system is not good but its not all bad it takes the votes away from the people,Its a winner take all system, and its taking people down who deserve to win even changing the senators to make them win. Its evil and makes no sense to what they are doing trying to give other parties a chance i get it but its gone a little to far.}

Read the following article: """# What Is the Electoral College? by the Office of the Federal Register

1 The Electoral College is a process, not a place. The founding fathers established it in the Constitution as a compromise between election of the President by a vote in Congress and election of the President by a popular vote of qualified citizens.

2 The Electoral College process consists of the selection of the electors, the meeting of the electors where they vote for President and Vice President, and the counting of the electoral votes by Congress.

3 The Electoral College consists of 538 electors. A majority of 270 electoral votes is required to elect the President. Your state’s entitled allotment of electors equals the number of members in its Congressional delegation: one for each member in the House of Representatives plus two for your Senators. . . .

4 Under the 23rd Amendment of the Constitution, the District of Columbia is allocated 3 electors and treated like a state for purposes of the Electoral College. For this reason, in the following discussion, the word “state” also refers to the District of Columbia.

5 Each candidate running for President in your state has his or her own group of electors. The electors are generally chosen by the candidate’s political party, but state laws vary on how the electors are selected and what their responsibilities are. . . .

6 The presidential election is held every four years on the Tuesday after the first Monday in November. You help choose your state’s electors when you vote for President because when you vote for your candidate you are actually voting for your candidate’s electors.

7 Most states have a “winner-take-all” system that awards all electors to the winning presidential candidate. However, Maine and Nebraska each have a variation of “proportional representation.” . . .

8 After the presidential election, your governor prepares a “Certificate of Ascertainment” listing all of the candidates who ran for President in your state along with the names of their respective electors. The Certificate of Ascertainment also declares the winning presidential candidate in your state and shows which electors will represent your state at the meeting of the electors in December of the election year. Your state’s Certificates of Ascertainments are sent to the Congress and the National Archives as part of the official records of the presidential election.

# The Indefensible Electoral College: Why even the best-laid defenses of the system are wrong by Bradford Plumer

9 What have Richard Nixon, Jimmy Carter, Bob Dole, the U.S. Chamber of Commerce, and the AFL-CIO all, in their time, agreed on? Answer: Abolishing the electoral college! They’re not alone; according to a Gallup poll in 2000, taken shortly after Al Gore—thanks to the quirks of the electoral college—won the popular vote but lost the presidency,1 over 60 percent of voters would prefer a direct election to the kind we have now. This year voters can expect another close election in which the popular vote winner could again lose the presidency. And yet, the electoral college still has its defenders. What gives? . . . What’s wrong with the electoral college

10 Under the electoral college system, voters vote not for the president, but for a slate of electors, who in turn elect the president. If you lived in Texas, for instance, and wanted to vote for John Kerry, you’d vote for a slate of 34 Democratic electors pledged to Kerry. On the offchance that those electors won the statewide election, they would go to Congress and Kerry would get 34 electoral votes. Who are the electors? They can be anyone not holding public office. Who picks the electors in the first place? It depends on the state. Sometimes state conventions, sometimes the state party’s central committee, sometimes the presidential candidates themselves. Can voters control whom their electors vote for? Not always. Do voters sometimes get confused about the electors and vote for the wrong candidate? Sometimes.

11 The single best argument against the electoral college is what we might call the disaster factor. The American people should consider themselves lucky that the 2000 fiasco was the biggest election crisis in a century; the system allows for much worse. Consider that state legislatures are technically responsible for picking electors, and that those electors could always defy the will of the people. Back in 1960, segregationists in the Louisiana legislature nearly succeeded in replacing the Democratic electors with new electors who would oppose John F. Kennedy. (So that a popular vote for Kennedy would not have actually gone to Kennedy.) In the same vein, “faithless” electors have occasionally refused to vote for their party’s candidate and cast a deciding vote for whomever they please. . . . Oh, and what if a state sends two slates of electors to Congress? It happened in Hawaii in 1960. Luckily, Vice President Richard Nixon, who was presiding over the Senate, validated only his opponent’s electors, but he made sure to do so “without establishing a precedent.” What if it happened again?

12 Perhaps most worrying is the prospect of a tie in the electoral vote. In that case, the election would be thrown to the House of Representatives, where state delegations vote on the president. (The Senate would choose the vice-president.) Because each state casts only one vote, the single representative from Wyoming, representing 500,000 voters, would have as much say as the 55 representatives from California, who represent 35 million voters. Given that many voters vote one party for president and another for Congress, the House’s selection can hardly be expected to reflect the will of the people. And if an electoral tie seems unlikely, consider this: In 1968, a shift of just 41,971 votes would have deadlocked the election; In 1976, a tie would have occurred if a mere 5,559 voters in Ohio and 3,687 voters in Hawaii had voted the other way. The election is only a few swing voters away from catastrophe.

13 At the most basic level, the electoral college is unfair to voters. Because of the winner-takeall system in each state, candidates don't spend time in states they know they have no chance of winning, focusing only on the tight races in the “swing” states. During the 2000 campaign, seventeen states didn’t see the candidates at all, including Rhode Island and South Carolina, and voters in 25 of the largest media markets didn’t get to see a single campaign ad. If anyone has a good argument for putting the fate of the presidency in the hands of a few swing voters in Ohio, they have yet to make it. . . .

14 It’s official: The electoral college is unfair, outdated, and irrational. The best arguments in favor of it are mostly assertions without much basis in reality. And the arguments against direct elections are spurious at best. It’s hard to say this, but Bob Dole was right: Abolish the electoral college!

# In Defense of the Electoral College: Five reasons to keep our despised method of choosing the President by Judge Richard A. Posner

15 The Electoral College is widely regarded as an anachronism,1 a non-democratic method of selecting a president that ought to be overruled by declaring the candidate who receives the most popular votes the winner. The advocates of this position are correct in arguing that the Electoral College method is not democratic in a modern sense . . . it is the electors who elect the president, not the people. When you vote for a presidential candidate you’re actually voting for a slate of electors.

16 But each party selects a slate of electors trusted to vote for the party’s nominee (and that trust is rarely betrayed) . . . however, it is entirely possible that the winner of the electoral vote will not win the national popular vote. Yet that has happened very rarely. It happened in 2000, when Gore had more popular votes than Bush yet fewer electoral votes, but that was the first time since 1888.

17 There are five reasons for retaining the Electoral College despite its lack of democratic pedigree;2 all are practical reasons, not liberal or conservative3 reasons.

## 1) Certainty of Outcome

18 A dispute over the outcome of an Electoral College vote is possible—--it happened in 2000--—but it’s less likely than a dispute over the popular vote. The reason is that the winning candidate’s share of the Electoral College invariably exceeds his share of the popular vote. In 2012’s election, for example, Obama4 received 61.7 percent of the electoral vote compared to only 51.3 percent of the popular votes cast for him and Romney.5 . . . Because almost all states award electoral votes on a winner-take-all basis, even a very slight plurality6 in a state creates a landslide electoral-vote victory in that state. A tie in the nationwide electoral vote is possible because the total number of votes—--538—--is an even number, but it is highly unlikely. . . .

## 2) Everyone’s President

19 The Electoral College requires a presidential candidate to have trans-regional appeal. No region (South, Northeast, etc.) has enough electoral votes to elect a president. So a solid regional favorite, such as Romney was in the South, has no incentive to campaign heavily in those states, for he gains no electoral votes by increasing his plurality in states that he knows he will win. This is a desirable result because a candidate with only regional appeal is unlikely to be a successful president. The residents of the other regions are likely to feel disenfranchised—to feel that their votes do not count, that the new president will have no regard for their interests, that he really isn’t their president.

## 3) Swing States

20 The winner-take-all method of awarding electoral votes induces the candidates—as we saw in 2012’s election—to focus their campaign efforts on the toss-up states . . . . Voters in toss-up states are more likely to pay close attention to the campaign—to really listen to the competing candidates—knowing that they are going to decide the election. They are likely to be the most thoughtful voters, on average (and for the further reason that they will have received the most information and attention from the candidates), and the most thoughtful voters should be the ones to decide the election.

## 4) Big States

21 The Electoral College restores some of the weight in the political balance that large states (by population) lose by virtue of the mal-apportionment of the Senate decreed in the Constitution. . . . The popular vote was very close in Florida in 2012; nevertheless Obama, who won that vote, got 29 electoral votes. A victory by the same margin in Wyoming would net the winner only 3 electoral votes. So, other things being equal, a large state gets more attention from presidential candidates in a campaign than a small state does. . . .

## 5) Avoid Run-Off Elections

22 The Electoral College avoids the problem of elections in which no candidate receives a majority of the votes cast. For example, Nixon in 1968 and Clinton in 1992 both had only a 43 percent plurality of the popular votes, while winning a majority in the Electoral College (301 and 370 electoral votes, respectively). There is pressure for run-off elections when no candidate wins a majority of the votes cast; that pressure, which would greatly complicate the presidential election process, is reduced by the Electoral College, which invariably produces a clear winner. . . .

23 It can be argued that the Electoral College method of selecting the president may turn off potential voters for a candidate who has no hope of carrying their state—Democrats in Texas, for example, or Republicans in California. Knowing their vote will have no effect, they have less incentive to pay attention to the campaign than they would have if the president were picked by popular vote . . . . But of course no voter’s vote swings a national election, and in spite of that, about one-half the eligible American population did vote in 2012’s election. Voters in presidential elections are people who want to express a political preference rather than people who think that a single vote may decide an election. . . ."""

Write a letter to your state senator in which you argue in favor of keeping the Electoral College or changing to election by popular vote for the president of the United States. Use the information from the texts in your essay. Manage your time carefully so that you can read the passages; plan your response; write your response; and revise and edit your response. Be sure to include a claim; address counterclaims; use evidence from multiple sources; and avoid overly relying on one source. Your response should be in the form of a multiparagraph essay. Write your response in the space provided.

text: {

## Appendix 2 - Post-Processing Script

import logging

import numpy as np

import polars as pl

import re

import textstat

import sys

from collections import Counter

from pathlib import Path

from sklearn.model\_selection import cross\_val\_predict

from sklearn.pipeline import FeatureUnion

from sklearn import metrics

from nltk import word\_tokenize, pos\_tag, PunktSentenceTokenizer

from sklearn.base import BaseEstimator, TransformerMixin

from xgboost import XGBClassifier

def trim\_to\_punctuation(input\_string):

match = re.search(r'[.!?;]([^.]\*)$', input\_string)

if match:

return input\_string[:match.start() + 1]

else:

return input\_string

def clean\_essays(essays):

r\_names = re.compile(r"\[Name\]|\[Your name\]|\[person\'s name\]", flags=re.ASCII)

r\_colon\_word\_start = re.compile(r"^\w+:\s\*", flags=re.MULTILINE)

r\_headings = re.compile('.\*:$', flags=re.MULTILINE)

r\_title\_line = re.compile(r'^(?!.\*[.?!])(\b\w+\b\s\*){0,10}$\n?', flags=re.MULTILINE)

r\_simple = re.compile(r"[^\w\s.,?\:!-'\"]+", flags=(re.ASCII|re.MULTILINE))

r\_newline = re.compile(r'(?<![.!?])\n', flags=(re.ASCII|re.MULTILINE))

r\_whitespace = re.compile(r"(\s)\1+", flags=(re.ASCII|re.MULTILINE))

essays = essays.to\_pandas()

essays['text'] = (

essays

.loc[:, 'text']

.str.replace('\\n', '\n')

.str.replace(r\_names, 'PROPER\_NAME', regex=True) # make datasynth match student

.str.replace(r\_colon\_word\_start, '', regex=True) # remove "Word:" line starts

.str.replace(r\_headings, '', regex=True) # remove "A Heading:"

.str.replace(r\_title\_line, '', regex=True) # remove "A Title or Heading" lines

.str.replace('#', '')

.str.replace('\*', '')

.str.replace('$', '')

.str.replace('\r', '')

.str.replace('\t', '')

.str.replace('\f ', '')

.str.replace('\v ', '')

.str.strip()

.apply(trim\_to\_punctuation)

.str.replace(r\_simple, '', regex=True) # only simple letters and numbers

.str.replace(r\_newline, ' ', regex=True) # replace random newline characters with space

.str.replace(r\_whitespace, r'\1', regex=True) # reduce repeated whitespace

.str.replace('\n', '\n\n')

)

essays = pl.from\_pandas(essays)

return essays

class CharacterCountTransformer(BaseEstimator, TransformerMixin):

def count\_characters(self, text):

return len(text)

def fit(self, X, y=None):

return self

def transform(self, X, y=None):

return np.array([self.count\_characters(text) for text in X]).reshape(-1, 1)

class SpecialCharCountTransformer(BaseEstimator, TransformerMixin):

def \_\_init\_\_(self):

self.special\_char\_re = re.compile(r'[@#$%^&\*():{}|<>]')

def count\_special\_chars(self, text):

return len(self.special\_char\_re.findall(text))

def fit(self, X, y=None):

return self

def transform(self, X, y=None):

return np.array([self.count\_special\_chars(text) for text in X]).reshape(-1, 1)

class SentenceCountTransformer(BaseEstimator, TransformerMixin):

def \_\_init\_\_(self):

self.tokenizer = PunktSentenceTokenizer()

def count\_sentences(self, text):

return len(self.tokenizer.sentences\_from\_text(text))

def fit(self, X, y=None):

return self

def transform(self, X, y=None):

return np.array([self.count\_sentences(text) for text in X]).reshape(-1, 1)

class WordCountTransformer(BaseEstimator, TransformerMixin):

def count\_words(self, text):

return len(word\_tokenize(text))

def fit(self, X, y=None):

return self

def transform(self, X, y=None):

return np.array([self.count\_words(text) for text in X]).reshape(-1, 1)

class FunctionWordFrequencyTransformer(BaseEstimator, TransformerMixin):

def calculate\_frequency(self, text):

function\_words = ['the', 'and', 'in', 'of', 'to']

return sum(text.lower().count(word) for word in function\_words)

def fit(self, X, y=None):

return self

def transform(self, X, y=None):

return np.array([self.calculate\_frequency(text) for text in X]).reshape(-1, 1)

class ReadabilityMetricsTransformer(BaseEstimator, TransformerMixin):

def calculate\_metrics(self, text):

return textstat.flesch\_reading\_ease(text), textstat.gunning\_fog(text)

def fit(self, X, y=None):

return self

def transform(self, X, y=None):

return np.array([self.calculate\_metrics(text) for text in X])

class POSTagsTransformer(BaseEstimator, TransformerMixin):

def pos\_tagging(self, text):

return pos\_tag(word\_tokenize(text))

def fit(self, X, y=None):

return self

def transform(self, X, y=None):

tags = ['NN', 'VB', 'JJ', 'RB']

pos\_counts = []

for text in X:

tag\_freq = Counter(tag for word, tag in self.pos\_tagging(text))

pos\_counts.append([tag\_freq.get(tag, 0) for tag in tags])

return np.array(pos\_counts)

def make\_featurizer():

character\_count\_transformer = CharacterCountTransformer()

special\_char\_count\_transformer = SpecialCharCountTransformer()

sentence\_count\_transformer = SentenceCountTransformer()

word\_count\_transformer = WordCountTransformer()

function\_word\_freq\_transformer = FunctionWordFrequencyTransformer()

readability\_metrics\_transformer = ReadabilityMetricsTransformer()

pos\_tags\_transformer = POSTagsTransformer()

combined\_features = FeatureUnion([

('character\_count', character\_count\_transformer),

('special\_char\_count', special\_char\_count\_transformer),

('sentence\_count', sentence\_count\_transformer),

('word\_count', word\_count\_transformer),

('function\_word\_freq', function\_word\_freq\_transformer),

('readability\_metrics', readability\_metrics\_transformer),

('pos\_tags', pos\_tags\_transformer)

], n\_jobs=-1)

feature\_columns = [ 'character\_count', 'special\_char\_count',

'sentence\_count', 'word\_count', 'function\_word\_freq',

'flesch\_reading\_ease', 'gunning\_fog', 'NN\_freq', 'VB\_freq',

'JJ\_freq', 'RB\_freq' ]

return combined\_features, feature\_columns

def compute\_features(essays):

featurizer, feature\_columns = make\_featurizer()

X\_transformed = featurizer.transform(essays.get\_column('text').to\_pandas())

essays\_transformed = pl.DataFrame(

X\_transformed,

schema=feature\_columns,

)

essays\_transformed = (

essays\_transformed

.select(essays.get\_column('id'), pl.all())

.join(essays, on='id')

)

return essays\_transformed

def filter\_with\_classifier(

df, target\_col, feature\_cols, classifier,

proba\_percentile=95, eval\_threshold=0.0005, max\_iterations=100, min\_eval=0.5,

):

prev\_metric = float('inf')

iteration = 0

while iteration < max\_iterations:

print(f"{iteration=}")

X = df.select(\*feature\_cols).to\_numpy()

y = df.get\_column(target\_col).to\_numpy()

y\_proba = cross\_val\_predict(classifier, X, y, cv=5, method='predict\_proba')

y\_proba = y\_proba[:, 1]

confidence\_threshold = np.percentile(y\_proba, proba\_percentile)

if confidence\_threshold <= 0.5:

break

high\_confidence\_indices = (y\_proba > confidence\_threshold)

correct\_classification\_indices = y\_proba.round() == y

to\_remove = high\_confidence\_indices & correct\_classification\_indices

df\_new = df.filter(~to\_remove)

print("row delta=", np.sum(to\_remove))

print("new rows=", len(df\_new))

X\_new = df\_new.select(\*feature\_cols).to\_numpy()

y\_new = df\_new.get\_column(target\_col).to\_numpy()

y\_proba = cross\_val\_predict(classifier, X\_new, y\_new, cv=5, method='predict\_proba')

y\_proba = y\_proba[:, 1]

new\_metric = metrics.roc\_auc\_score(y\_new, y\_proba)

print(f"{new\_metric=}")

print("metric delta=", prev\_metric - new\_metric)

if new\_metric < min\_eval:

break

if prev\_metric - new\_metric > eval\_threshold:

prev\_metric = new\_metric

df = df\_new

else:

break

iteration += 1

print()

return df, new\_metric

real\_essays = # load student essays

essays\_transformed = compute\_features(real\_essays)

clf = XGBClassifier(random\_state=0)

filter\_cols = ['character\_count', 'word\_count', 'sentence\_count', 'function\_word\_freq']

essays\_filtered, final\_score = filter\_with\_classifier(

essays\_transformed,

target\_col='generated',

feature\_cols=filter\_cols,

classifier=clf,

proba\_percentile=95,

eval\_threshold=-1,

max\_iterations=100,

min\_eval=0.535,

)

print(f"{final\_score=}")