In [2]:

```
import pandas as pd

# load
df = pd.read_csv('balanced_ai_human_prompts.csv')
df.head()
```

Out[2]:

	text	generated
0	Machine learning, a subset of artificial intel...	1
1	A decision tree, a prominent machine learning ...	1
2	Education, a cornerstone of societal progress,...	1
3	Computers, the backbone of modern technology, ...	1
4	Chess, a timeless game of strategy and intelle...	1

In [3]:

```
# check mean length in text column, vocabulary size in text colomn, and  num
mean_length = df['text'].apply(len).mean()
vocab_size = len(set(' '.join(df['text']).split()))
num_entries = len(df)
print(f'Mean length: {mean_length}')
print(f'Vocabulary size: {vocab_size}')
print(f'Number of entries: {num_entries}')
```

```
Mean length: 1670.9229090909091
Vocabulary size: 31788
Number of entries: 2750
```