Dummy Baseline Model

This notebook creates a baseline model by extracting the first 1000 characters of the abstracts.

In [ ]:
```
import pandas as pd
```

In [2]:
```
from google.colab import drive

drive.mount("/content/drive")
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, cal
l drive.mount("/content/drive", force_remount=True).
```

Load Validation Data

In [3]:
```
INPUT_FILE_DIR = "/content/drive/MyDrive/BioLaySumm2024_main/extracted/"
ELIFE_VAL_FILE_NAME = "eLife_val_extracted.csv"
PLOS_VAL_FILE_NAME = "PLOS_val_extracted.csv"

val_eLife_df = pd.read_csv(INPUT_FILE_DIR + ELIFE_VAL_FILE_NAME)

val_plos_df = pd.read_csv(INPUT_FILE_DIR + PLOS_VAL_FILE_NAME)
```

In [6]:
```
val_eLife_df['abstract'].apply(lambda x: x[:1000])
```

Out[6]:
```
0      Cell-fate reprograming is at the heart of deve...
1      Klebsiella pneumoniae is a respiratory , blood...
2      Plasmodium vivax relapse infections occur foll...
3      When 2 Mha of Amazonian forests are disturbed ...
4      Neurosphere formation is commonly used as a su...
                             ...
236    Similarity between two individuals in the comb...
237    Early development of many animals is character...
238    Gene control systems sometimes interpret multi...
239    The number of neurotransmitter-filled vesicles...
240    HIV-1 Gag selects and packages a dimeric , uns...
Name: abstract, Length: 241, dtype: object
```

Output Validation data

In [19]:
```
OUTPUT_FILE_DIR = "/content/drive/MyDrive/BioLaySumm2024-evaluation_scripts/

val_eLife_df['abstract'].apply(lambda x: x[:1000]).to_csv(OUTPUT_FILE_DIR+'e

val_plos_df['abstract'].apply(lambda x: x[:1000]).to_csv(OUTPUT_FILE_DIR+'pl
```

Load test data

In [7]:
```
ELIFE_TEST_FILE_NAME = "eLife_test_extracted.csv"
PLOS_TEST_FILE_NAME = "PLOS_test_extracted.csv"


test_eLife_df = pd.read_csv(INPUT_FILE_DIR + ELIFE_TEST_FILE_NAME)

test_plos_df = pd.read_csv(INPUT_FILE_DIR + PLOS_TEST_FILE_NAME)
```

In [8]:
```
test_eLife_df
```

Out[8]:

	Unnamed: 0	headings	keywords	id	abstract
0	0	['Abstract', 'Introduction', 'Results and disc...	['biochemistry and chemical biology', 'computa...	elife-81547-v1	Acylation of diverse carbohydrates occurs acro...
1	1	['Abstract', 'Introduction', 'Results', 'Discu...	['computational and systems biology']	elife-86176-v2	Honey bee ecology demands they make both rapid...
2	2	['Abstract', 'Introduction', 'Results', 'Discu...	['genetics and genomics']	elife-82210-v1	Biguanides , including the world’s most prescr...
3	3	['Abstract', 'Introduction', 'Results', 'Discu...	['microbiology and infectious disease', 'ecolo...	elife-83152-v2	Ecological relationships between bacteria medi...
4	4	['Abstract', 'Introduction', 'Results', 'Discu...	['neuroscience']	elife-83044-v2	Gamma oscillations are believed to underlie co...
...	...	...	...	...	...
137	137	['Abstract', 'Introduction', 'Results', 'Discu...	['tools and resources', 'genetics and genomics']	elife-84831-v1	High-throughput transgenesis using synthetic D...
138	138	['Abstract', 'Introduction', 'Results', 'Discu...	['structural biology and molecular biophysics'...	elife-82885-v1	To reach their final destinations , outer memb...
139	139	['Abstract', 'Introduction', 'Results', 'Discu...	['neuroscience']	elife-87902-v2	Midbrain dopamine (DA) neurons are key regul...
140	140	['Abstract', 'Introduction', 'Methods', 'Resul...	['medicine', 'epidemiology and global health']	elife-79615-v2	Mobile health (mHealth) interventions , whic...
141	141	['Abstract', 'Introduction', 'Methods', 'Resul...	['medicine', 'epidemiology and global health']	elife-80428-v2	Severe acute respiratory syndrome coronavirus ...

142 rows × 5 columns

Output Test Data

In [16]:
```
OUTPUT_FILE_DIR = "/content/drive/MyDrive/BioLaySumm2024_main/test_submissio

test_eLife_df['abstract'].apply(lambda x: x[:1000]).to_csv(OUTPUT_FILE_DIR+'

test_plos_df['abstract'].apply(lambda x: x[:1000]).to_csv(OUTPUT_FILE_DIR+'p
```

In [ ]: