Create mini dataset for Evaluation

Because running evaluation code on full eval datasets is very long (>12hrs), so we can create a mini version of the dataset (with about 10% randomly picked data)

In [1]:

```
import pandas as pd
```

In [2]:

```
filepath = "./data/biolaysumm2024_data/"
filename = "eLife_val.jsonl"
```

In [3]:

```
df = pd.read_json(filepath + filename,
                  orient="records",
                  lines=True
                 )
df.head()
```

Out[3]:

	lay_summary	article	headings	keywords	id
0	The DNA in genes encodes the basic information...	Cell-fate reprograming is at the heart of deve...	[Abstract, Introduction, Results, Discussion, ...	[developmental biology]	elife-15477-v3
1	Klebsiella pneumoniae is a type of bacteria th...	Klebsiella pneumoniae is a respiratory , blood...	[Abstract, Introduction, Results, Discussion, ...	[microbiology and infectious disease, immunolo...	elife-56656-v2
2	Malaria is one of the world's most deadly infe...	Plasmodium vivax relapse infections occur foll...	[Abstract, Introduction, Results, Discussion, ...	[epidemiology and global health]	elife-04692-v2
3	The Amazon rainforest in South America is the ...	When 2 Mha of Amazonian forests are disturbed ...	[Abstract, Introduction, Results, Discussion, ...	[ecology]	elife-21394-v2
4	Neurons that arise in the adult nervous system...	Neurosphere formation is commonly used as a su...	[Abstract, Introduction, Results, Discussion, ...	[stem cells and regenerative medicine]	elife-02669-v2

In [4]:

```
n = len(df)
print("Number of rows =", n)
```

```
Number of rows = 241
```

In [5]:

```
ratio = 0.1 # 10%

mini_df = df.sample(frac=ratio,
                    random_state=42
                   )
```

```python
mini_df.head()
```

Out[5]:

	lay_summary	article	headings	keywords	id
24	It can take several months , or even years , f...	Mature neural networks synchronize and integra...	[Abstract, Introduction, Results, Discussion, ...	[neuroscience]	elife-69011-v2
6	Many of our decisions are made on the basis of...	Many decisions are thought to arise via the ac...	[Abstract, Introduction, Results, Discussion, ...	[neuroscience]	elife-17688-v1
222	Oculo-Cerebro-Renal syndrome of Lowe (Lowe sy...	Mutations in the inositol 5-phosphatase OCRL c...	[Abstract, Introduction, Results, Discussion, ...	[cell biology]	elife-02975-v2
208	When an embryo develops , its cells must work ...	Gradients of signaling proteins are essential ...	[Abstract, Introduction, Results, Discussion, ...	[developmental biology]	elife-38137-v3
236	Our genomes contain a record of historical eve...	Similarity between two individuals in the comb...	[Abstract, Introduction, Results, Discussion, ...	[evolutionary biology, genetics and genomics]	elife-15266-v1

In [6]:

```python
print("Number of rows =", len(mini_df))
print("Ratio =", len(mini_df) / len(df))
```

```
Number of rows = 24
Ratio = 0.0995850622406639
```

In [7]:

```python
output_path = "./data/mini_dataset/"
output_file = "eLife_val_mini.jsonl"

print("Writing mini set to", output_file)
mini_df.to_json(output_path + output_file,
                orient="records",
                lines=True
               )
print("Completed")
```

```
Writing mini set to eLife_val_mini.jsonl
Completed
```

In [8]:

```python
# put the code into a function, so we can apply it to multiple dataset
def make_mini_dataset(full_input_filename, full_output_filename, ratio=0.1):
    """
        read data from a full set, pick a random sample and write to output file
    """
    print("reading from file =", full_input_filename)

    df = pd.read_json(full_input_filename,
                      orient="records",
                      lines=True
                     )
    # print(df.head())
    print("Number of records =", len(df))


    mini_df = df.sample(frac=ratio,
                        random_state=42
                       )

    # print(mini_df.head())
    print("Number of rows in mini set =", len(mini_df))

    print("Writing mini set to", output_file)
    mini_df.to_json(full_output_filename,
                    orient="records",
                    lines=True
                   )
    print("Completed")
```

In [9]:
```
make_mini_dataset("./data/biolaysumm2024_data/eLife_val.jsonl",
                  "./data/mini_dataset/eLife_val_mini.jsonl")
```

```
reading from file = ./data/biolaysumm2024_data/eLife_val.jsonl
Number of records = 241
Number of rows in mini set = 24
Writing mini set to eLife_val_mini.jsonl
Completed
```

In [10]:
```
make_mini_dataset("./data/biolaysumm2024_data/PLOS_val.jsonl",
                  "./data/mini_dataset/PLOS_val_mini.jsonl")
```

```
reading from file = ./data/biolaysumm2024_data/PLOS_val.jsonl
Number of records = 1376
Number of rows in mini set = 138
Writing mini set to eLife_val_mini.jsonl
Completed
```