Data Extraction for Mini dataset

In this notebook, we will extract abstract and summary parts from articles in evaluation, then write to csv files

In [1]:
```
import pandas as pd
```

In [2]:
```
# function to extract first paragraph of the text
def get_abstract(text, max_char = 1500):
    """
        return abstract (first paragraph) of the text
    """
    result = ""
    result = text.split("\n")[0][:max_char]
    return result
```

Apply data extraction for validation datasets

In [3]:
```
file_path = "./data/mini_dataset/"
file_names = ["eLife_val_mini.jsonl",
              "PLOS_val_mini.jsonl"
             ]

output_path = "./data/mini_dataset/"

print("Abstract text extraction:")
print("============================")
for filename in file_names:
    print("Processing file =", filename)
    df = pd.read_json(file_path+filename,
                      orient="records",
                      lines=True)
    print("Number of records =", len(df))

    # apply get_abstract function
    print("Getting abstracts")
    df["abstract"] = df["article"].apply(get_abstract)
    print("Completed")

    output_df = df["abstract"]
    print(output_df.head())
    output_filename = filename[:filename.rfind(".")] + "_extracted.txt"
    print("Writing output to", output_filename)
    output_df.to_csv(output_path + output_filename,
                     index=False,
                     header=False,
                     sep="\n"
                    )
    print("Completed")
    print("--------------------")
```

```
print("======= completed ========")
```

```
Abstract text extraction:
============================
Processing file = eLife_val_mini.jsonl
Number of records = 24
Getting abstracts
Completed
0    Mature neural networks synchronize and integra...
1    Many decisions are thought to arise via the ac...
2    Mutations in the inositol 5-phosphatase OCRL c...
3    Gradients of signaling proteins are essential ...
4    Similarity between two individuals in the comb...
Name: abstract, dtype: object
Writing output to eLife_val_mini_extracted.txt
Completed
--------------------
Processing file = PLOS_val_mini.jsonl
Number of records = 138
Getting abstracts
Completed
0    Fleas can transmit Yersinia pestis by two mech...
1    Endogenous retroviruses ( ERVs ) are remnants ...
2    The Drosophila embryonic gonad is assembled fr...
3    Recently , we presented a study of adult neuro...
4    Understanding the transcriptional regulation o...
Name: abstract, dtype: object
Writing output to PLOS_val_mini_extracted.txt
Completed
--------------------
======= completed ========
```

In [ ]: