AI IN COVID LITERATURE ANALYSIS

National Computational Infrastructure

Label Set

- 2 levels of labels
- Each second-level label belongs to a first-level label

Model Training

- Finetuning **Google's Flan-t5** (encoder-decoder)
- **Conditional Loss** – masking out sub-labels belonging to unlikely large labels.
- Prepend a task name to the model input.

```
def inputForT5(t:str, a:str):
    return f'infer academic subject. title: {t}. abstract: {a}'
```

- **Extra training epochs** for infrequent labels to alleviate under-training due to Conditional Loss

infer academic subject from title.
title: COVID-19 and dysnatremia: A comparison between COVID-19 and non-COVID-19 respiratory illness

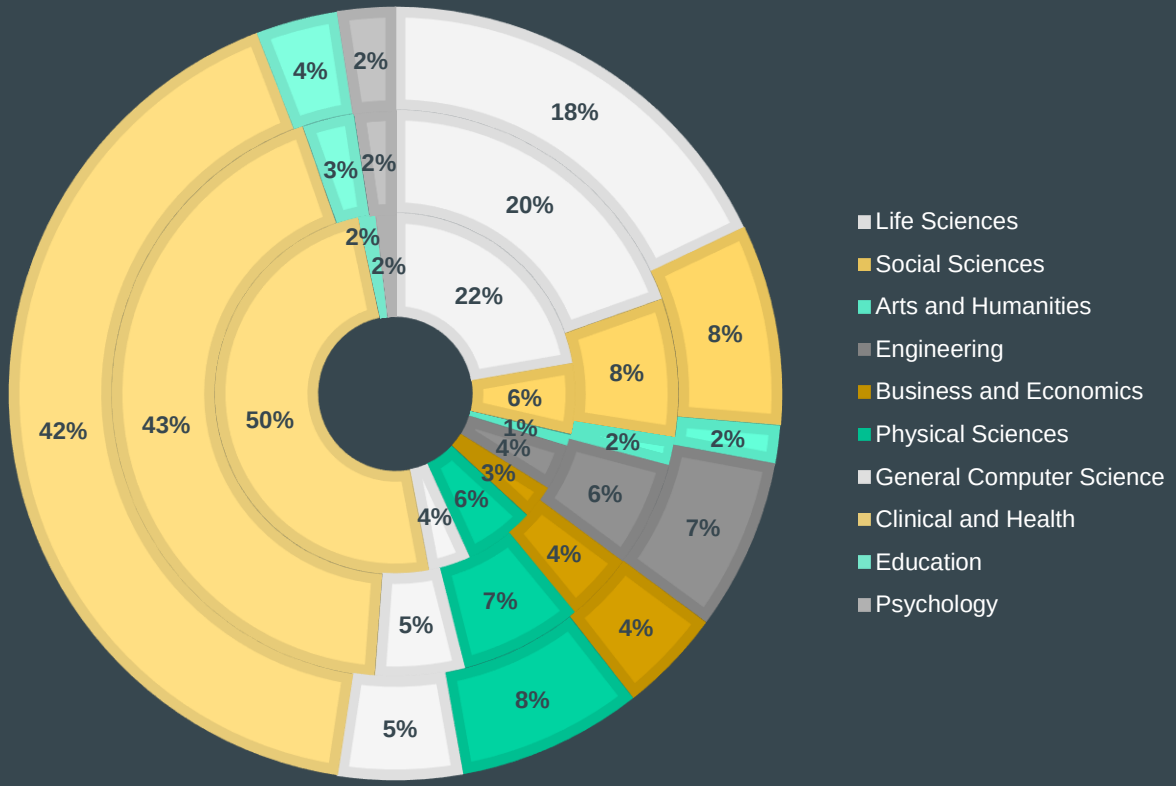
Compute ctrl+Enter 0.3

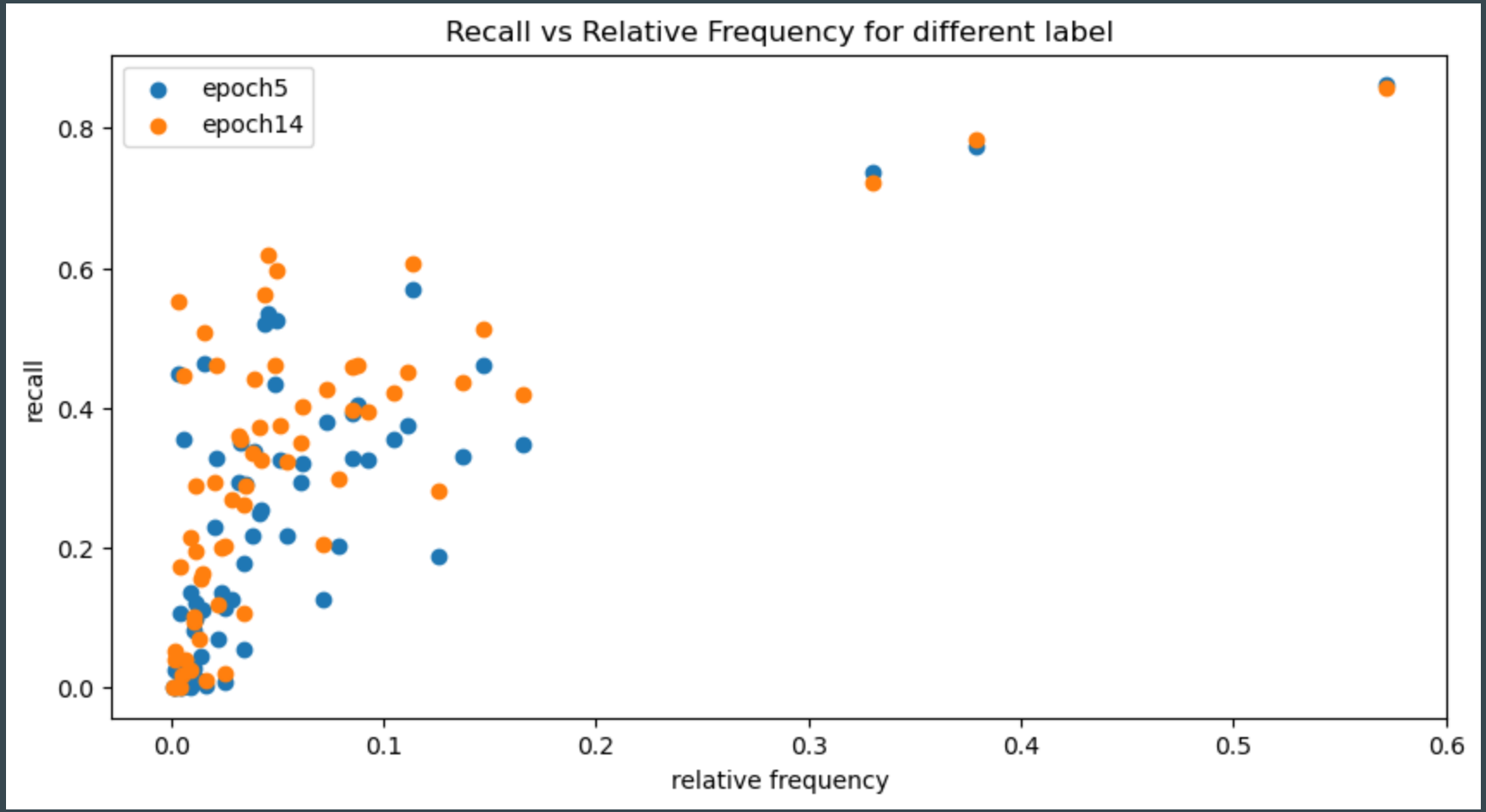
Computation time on Intel Xeon 3rd Gen Scalable cpu: 0.095 s

Science

Key Findings

- **Medical & Biological** topics contribute the most (as expected)
- Percentage of **other topics grows** with time




Recall vs Relative Frequency for different label
epoch5
epoch14
recall
relative frequency
0.0
0.2
0.4
0.6
0.8
0.0
0.1
0.2
0.3
0.4
0.5
0.6