Masked Language Modeling YouTube Captioning Errors?

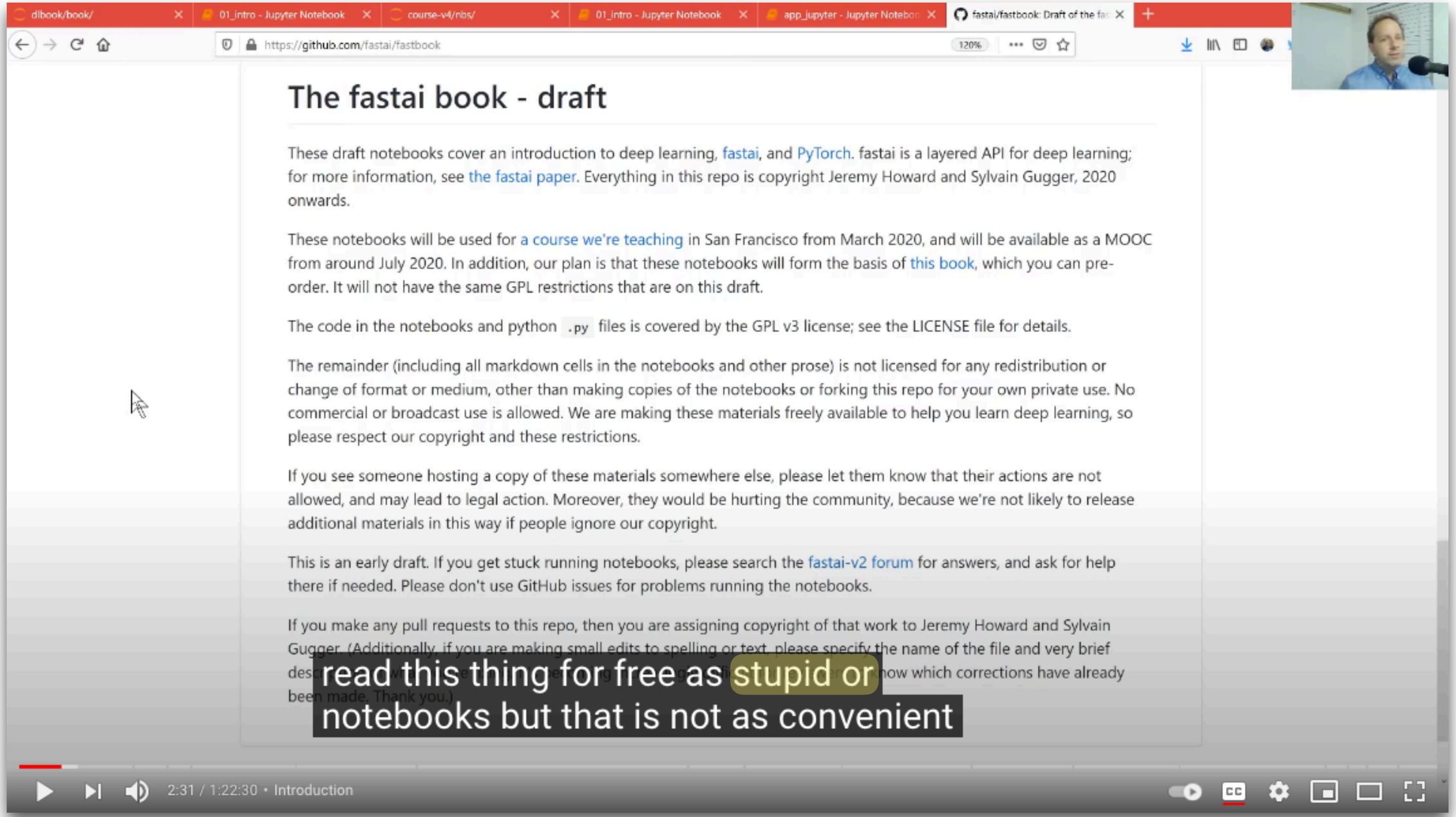
Emily McMilin

Masked Language Modeling YouTube Captioning Errors?

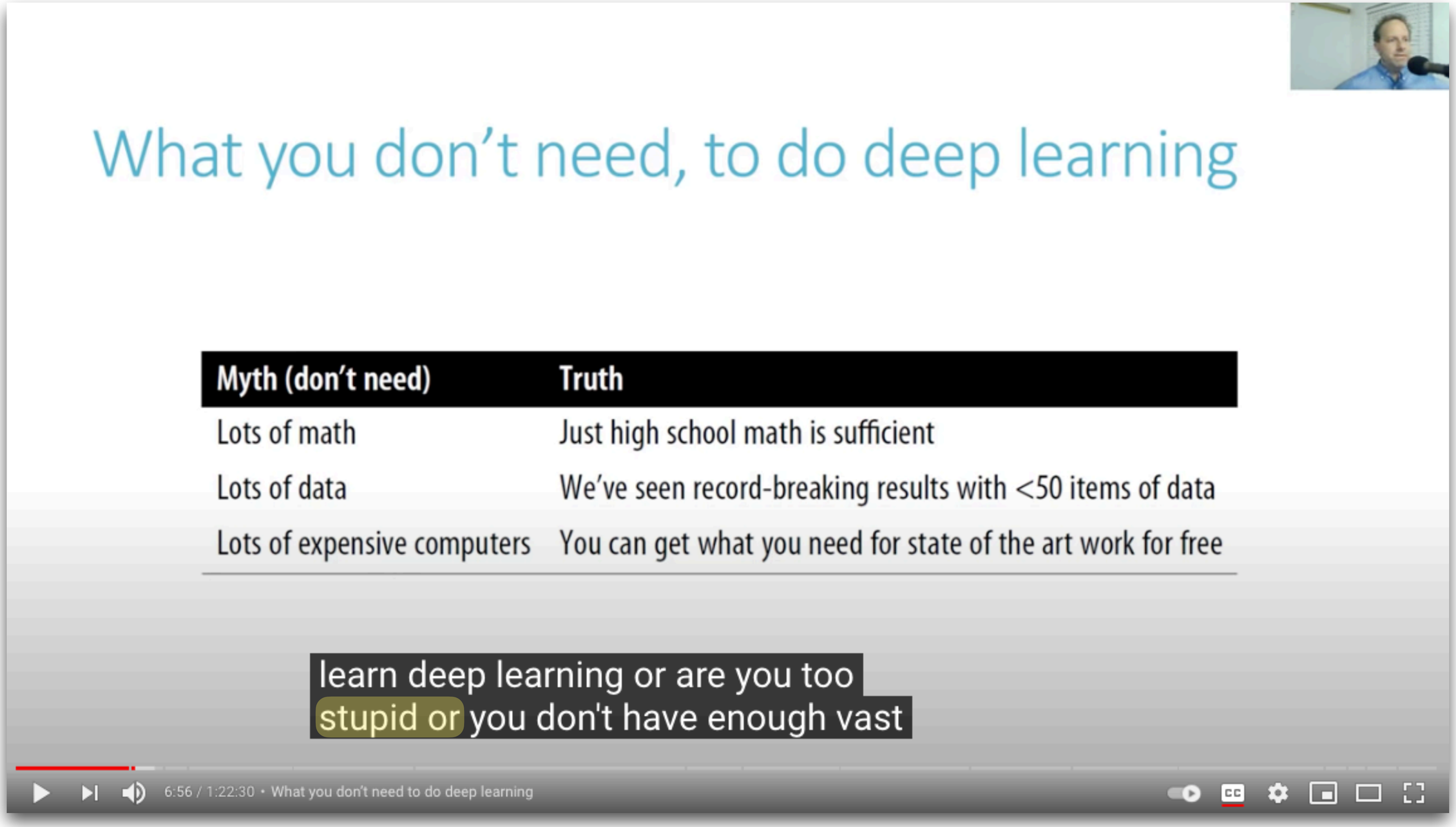
Emily McMilin

Motivation

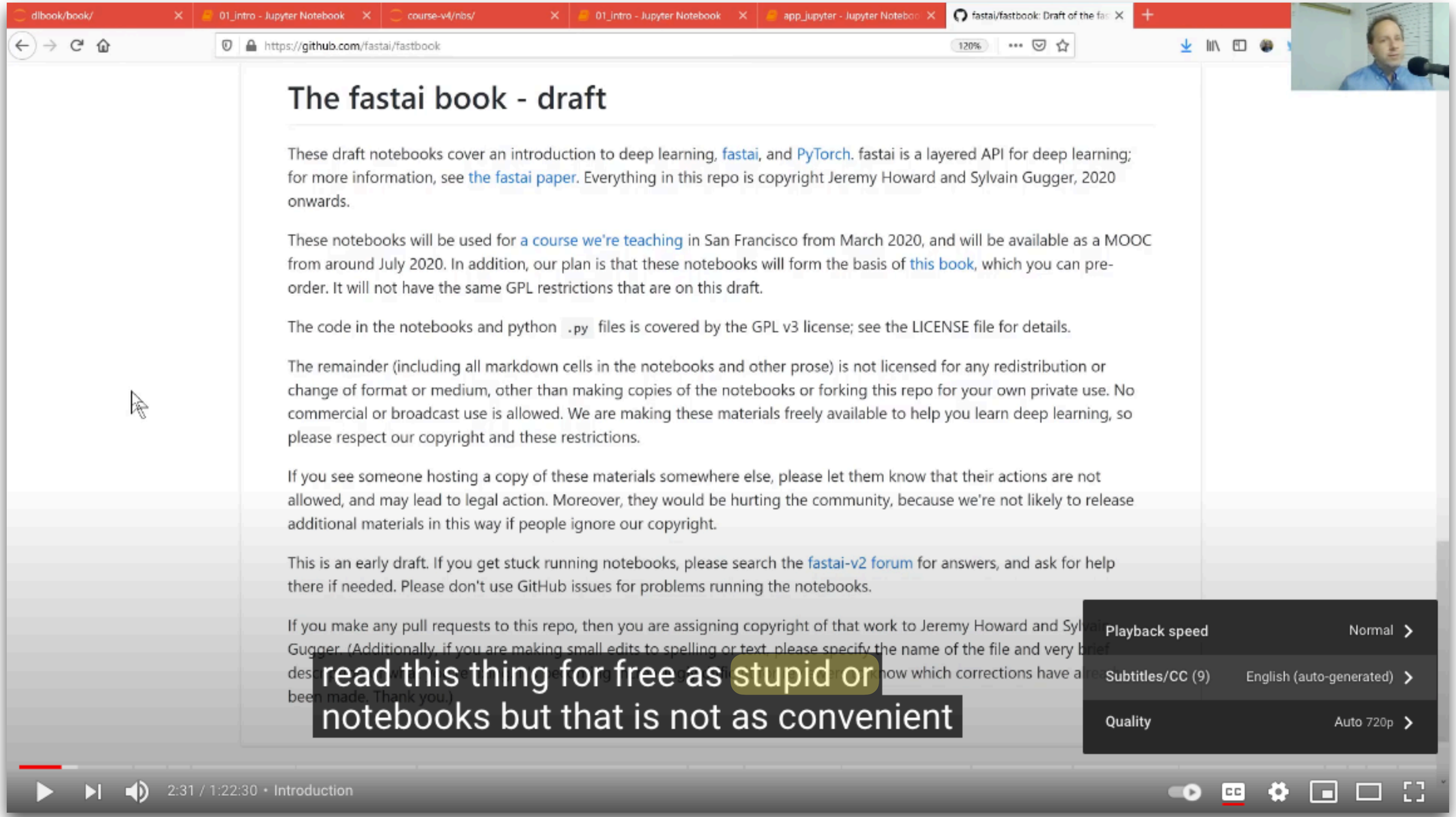
Motivation: stupid or...?



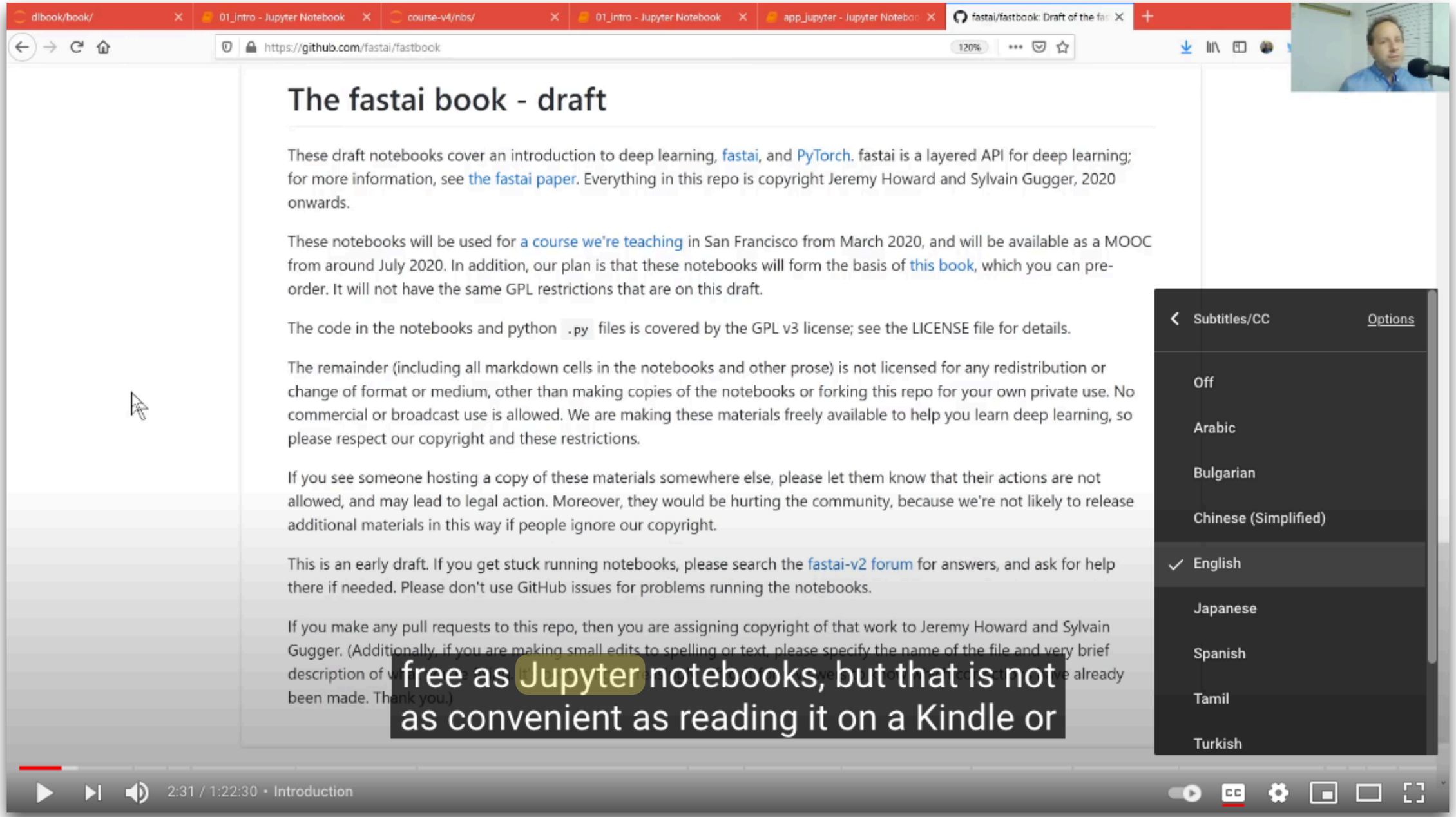
Motivation: stupid or...?



Motivation: This MLM needs a label

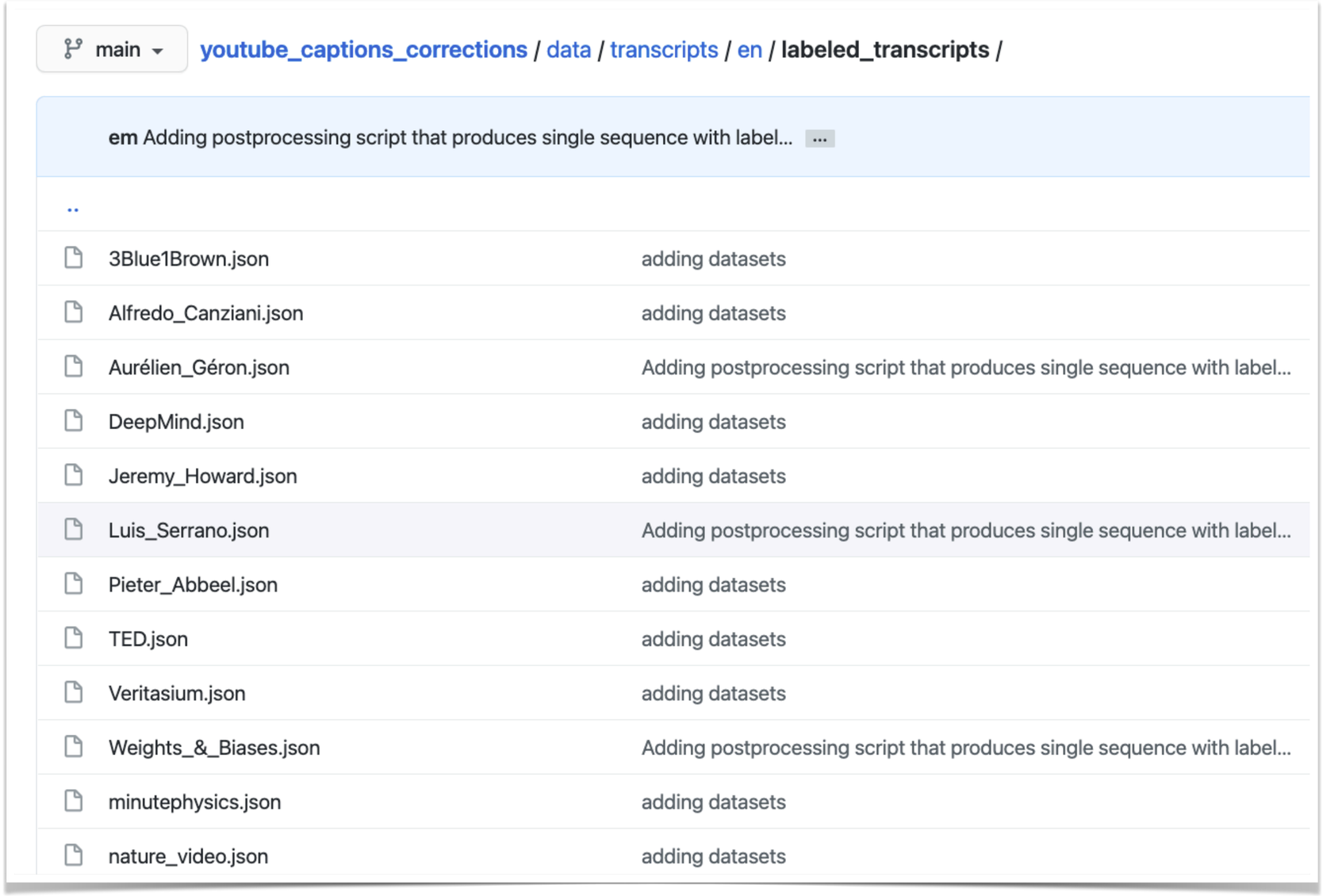


Motivation: Label provided!



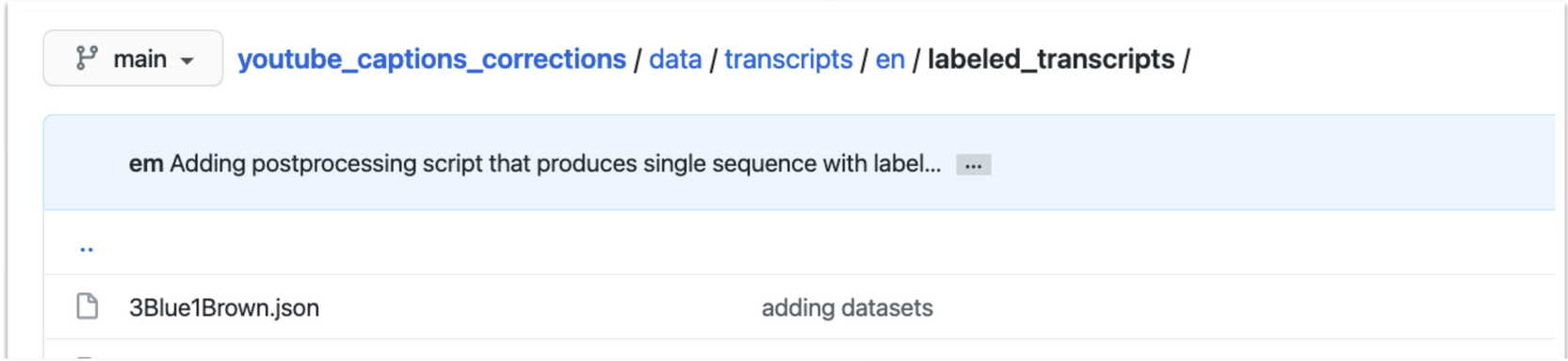
Dataset

Dataset: Corrections from my fav creators



https://github.com/2dot71mily/youtube_captions_corrections

Dataset: Diff tools!



class `difflib.`**`Differ`**

This is a class for comparing sequences of lines of text, and producing human-readable differences or deltas. Differ uses `SequenceMatcher` both to compare sequences of lines, and to compare sequences of characters within similar (near-matching) lines.

Each line of a `Differ` delta begins with a two-letter code:

Code	Meaning
`'- '`	line unique to sequence 1
`'+ '`	line unique to sequence 2
`'  '`	line common to both sequences

https://docs.python.org/3/library/difflib.html

Dataset: Jeremy Howard example

Video metadata

In [4]:
```
transcripts[['video_titles', 'playlist_ids', 'channel_ids']]
```

Out[4]:

	video_titles	playlist_ids	channel_ids
5L3Ao5KuCC4	Lesson 3 - Deep Learning for Coders (2020)	PLfYUBJiXbdtRL3FMB3GoWHRI8ieU6FhfM	UCX7Y2qWriXpqocG97SFW2OQ
BvHmRx14HQ8	Lesson 2 - Deep Learning for Coders (2020)	PLfYUBJiXbdtRL3FMB3GoWHRI8ieU6FhfM	UCX7Y2qWriXpqocG97SFW2OQ
CzdWqFTmn0Y	Intro to Machine Learning: Lesson 1	PLfYUBJiXbdtSyktd8A_x0JNd6lxDcZE96	UCX7Y2qWriXpqocG97SFW2OQ
IPBSB1HLNLo	Lesson 1: Deep Learning 2018	PLfYUBJiXbdtS2UQRzyrxmyVHoGW0gmLSM	UCX7Y2qWriXpqocG97SFW2OQ
JNxcznsrRb8	Lesson 2: Deep Learning 2018	PLfYUBJiXbdtS2UQRzyrxmyVHoGW0gmLSM	UCX7Y2qWriXpqocG97SFW2OQ
VEG5xT5gAHc	Lesson 7 - Deep Learning for Coders (2020)	PLfYUBJiXbdtRL3FMB3GoWHRI8ieU6FhfM	UCX7Y2qWriXpqocG97SFW2OQ
WjnwWeGjZcM	Lesson 8 - Deep Learning for Coders (2020)	PLfYUBJiXbdtRL3FMB3GoWHRI8ieU6FhfM	UCX7Y2qWriXpqocG97SFW2OQ
XfoYk_Z5AkI	Lesson 1: Deep Learning 2019 - Image classific...	PLfYUBJiXbdtSIJb-Qd3pw0cqCbkGeS0xn	UCX7Y2qWriXpqocG97SFW2OQ
_QUEXsHfsA0	Lesson 1 - Deep Learning for Coders (2020)	PLfYUBJiXbdtRL3FMB3GoWHRI8ieU6FhfM	UCX7Y2qWriXpqocG97SFW2OQ
cX30jxMNBUw	Lesson 6 - Deep Learning for Coders (2020)	PLfYUBJiXbdtRL3FMB3GoWHRI8ieU6FhfM	UCX7Y2qWriXpqocG97SFW2OQ
ccMHJeQU4Qw	Lesson 2: Deep Learning 2019 - Data cleaning a...	PLfYUBJiXbdtSIJb-Qd3pw0cqCbkGeS0xn	UCX7Y2qWriXpqocG97SFW2OQ
krIVOb23EH8	Lesson 5 - Deep Learning for Coders (2020)	PLfYUBJiXbdtRL3FMB3GoWHRI8ieU6FhfM	UCX7Y2qWriXpqocG97SFW2OQ
p50s63nPq9I	Lesson 4 - Deep Learning for Coders (2020)	PLfYUBJiXbdtRL3FMB3GoWHRI8ieU6FhfM	UCX7Y2qWriXpqocG97SFW2OQ

https://github.com/2dot71mily/youtube_captions_corrections

Dataset: Jeremy Howard example

Transcript text and metadata

```
In [5]: transcripts[['autogen', 'manual','autogen_text', 'manual_text', 'diffs',]]
```

Out[5]:

	autogen	manual	autogen_text	manual_text	diffs
5L3Ao5KuCC4	[{'text': 'oh hello and welcome to lesson thre...	[{'text': 'So hello, and welcome to Lesson 3 o...	oh hello and welcome to lesson three of practi...	So hello, and welcome to Lesson 3 of Practical...	[{'-': ['oh', 'hello'], '+': ['So', 'hello,']}...
BvHmRx14HQ8	[{'text': 'so uh hello everybody and welcome b...	[{'text': 'So, hello everybody, and welcome ba...	so uh hello everybody and welcome back to prac...	So, hello everybody, and welcome back to Pract...	[{'-': ['so', 'uh'], '+': ['So,']}, hello, {'-...
CzdWqFTmn0Y	[{'text': 'Oh', 'start': 0.0, 'duration': 9.05...	[{'text': 'Okay, so let me introduce everybody...	Oh good okay so let me introduce everybody to ...	Okay, so let me introduce everybody to everybo...	[{'-': ['Oh', 'good', 'okay'], '+': ['Okay,']}...
IPBSB1HLNLo	[{'text': 'hi everybody welcome to practical d...	[{'text': 'Hi everybody welcome to practical d...	hi everybody welcome to practical deep learnin...	Hi everybody welcome to practical deep learnin...	[{'-': ['hi'], '+': ['Hi']}, everybody, welcom...
JNxcznsrRb8	[{'text': 'okay so welcome back to deep learni...	[{'text': 'Okay so welcome back to deep learni...	okay so welcome back to deep learning lesson 2...	Okay so welcome back to deep learning lesson 2...	[{'-': ['okay'], '+': ['Okay']}, so, welcome, ...
VEG5xT5gAHc	[{'text': 'hi everybody and welcome to lesson ...	[{'text': 'Hi everybody and welcome to lesson ...	hi everybody and welcome to lesson 7 we're goi...	Hi everybody and welcome to lesson 7! We're\ng...	[{'-': ['hi'], '+': ['Hi']}, everybody, and, w...
WjnwWeGjZcM	[{'text': 'hi everybody and welcome to lesson ...	[{'text': 'Hi everybody and welcome to lesson ...	hi everybody and welcome to lesson eight the l...	Hi everybody and welcome to lesson eight,\nthe...	[{'-': ['hi'], '+': ['Hi']}, everybody, and, w...
XfoYk_Z5AkI	[{'text': 'okay so welcome practical deep lear...	[{'text': 'Okay', 'start': 0.159, 'duration': ...	okay so welcome practical deep learning for co...	Okay so Welcome Practical deep learning for co...	[{'-': ['okay'], '+': ['Okay']}, so, {'-': ['w...
_QUEXsHfsA0	[{'text': 'so hello everybody and welcome to d...	[{'text': 'So hello everybody and welcome to D...	so hello everybody and welcome to deep learnin...	So hello everybody and welcome to Deep Learnin...	[{'-': ['so'], '+': ['So']}, hello, everybody,...
cX30jxMNBUw	[{'text': 'hi everybody and welcome to lesson ...	[{'text': 'Hi everybody and welcome to Lesson ...	hi everybody and welcome to lesson six where w...	Hi everybody and welcome to Lesson 6, where\nw...	[{'-': ['hi'], '+': ['Hi']}, everybody, and, w...
ccMHJeQU4Qw	[{'text': 'welcome to lesson two where we're g...	[{'text': 'Welcome to Lesson 2 where we're goi...	welcome to lesson two where we're going to be ...	Welcome to Lesson 2 where we're going to be ta...	[{'-': ['welcome'], '+': ['Welcome']}, to, {'-...
krIVOb23EH8	[{'text': 'welcome to lesson five and we'll be...	[{'text': 'Welcome to lesson five and we'll be...	welcome to lesson five and we'll be talking ab...	Welcome to lesson five and we'll be talking\na...	[{'-': ['welcome'], '+': ['Welcome']}, to, les...
p50s63nPq9I	[{'text': 'welcome back and here is lesson or ...	[{'text': 'Welcome back and here is lesson 4 w...	welcome back and here is lesson or which is wh...	Welcome back and here is lesson 4 which is\nwh...	[{'-': ['welcome'], '+': ['Welcome']}, back, a...

https://github.com/2dot71mily/youtube_captions_corrections

Dataset: Jeremy Howard example

Label data

In [42]:
```
transcripts[['is_autogen_unique', 'autogen_seq', 'manual_seq', 'manual_addl_rep']]
```

Out[42]:

	is_autogen_unique	autogen_seq	manual_seq	manual_addl_rep
5L3Ao5KuCC4	[2, 2, 0, 0, 0, 2, 2, 0, 2, 2, 2, 0, 2, 2, 0, ...	[oh, hello, , , , lesson, three, , practical, ...	[So hello,, So hello,, , , , Lesson 3, Lesson ...	[1, 1, 0, 0, 0, 1, 1, 0, 2, 2, 2, 0, 1, 1, 0, ...
BvHmRx14HQ8	[2, 2, 0, 2, 0, 0, 0, 0, 2, 2, 2, 0, 2, 2, 0, ...	[so, uh, , everybody, , , , , practical, deep,...	[So,, So,, , everybody,, , , , , Practical Dee...	[1, 1, 0, 0, 0, 0, 0, 0, 2, 2, 2, 0, 1, 1, 0, ...
CzdWqFTmn0Y	[2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 2, ...	[Oh, good, okay, , , , , , , , else, , , , so,...	[Okay,, Okay,, Okay,, , , , , , , , else,, , ,...	[2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
IPBSB1HLNLo	[2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[hi, , , , , , , , , , , , , , , , course, , p...	[Hi, , , , , , , , , , , , , , , , course,, , ...	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
JNxcznsrRb8	[2, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 2, 0, 0, ...	[okay, , , , , , , , , last, , , got, , , , , ...	[Okay, , , , , , , , , Last, , , Got, , , , , ...	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
VEG5xT5gAHc	[2, 0, 0, 0, 0, 0, 0, 2, 2, 0, 0, 0, 0, 0, 0, ...	[hi, , , , , , 7, we're, , , , , , , , , , , ,...	[Hi, , , , , , 7! We're, 7! We're, , , , , , ,...	[0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, ...
WjnwWeGjZcM	[2, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, ...	[hi, , , , , , eight, , , , , , , , , course, ...	[Hi, , , , , , eight,, , , , , , , , course....	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
XfoYk_Z5AkI	[2, 0, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[okay, , welcome, practical, , , , , , , , , , ,...	[Okay, , Welcome Practical, Welcome Practical,...	[0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
_QUEXsHfsA0	[2, 0, 0, 0, 0, 0, 0, 2, 2, 0, 2, 2, 2, 2, 0, ...	[so, , , , , , deep, learning, , coders, lesso...	[So, , , , , , Deep Learning, Deep Learning, ,...	[0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 3, 3, 3, 3, 0, ...
cX30jxMNBUw	[2, 0, 0, 0, 0, 0, 2, 2, 0, 0, 0, 0, 0, 0, 0, ...	[hi, , , , , lesson, six, , , , , , , , , , ...	[Hi, , , , , Lesson 6,, Lesson 6,, , , , , , ,...	[0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, ...
ccMHJeQU4Qw	[2, 0, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[welcome, , lesson, two, , , , , , , , , , , ,...	[Welcome, , Lesson 2, Lesson 2, , , , , , , , ...	[0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
krlVOb23EH8	[2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[welcome, , , , , , , , , , , , , , , , , , , ...	[Welcome, , , , , , , , , , , , , , , , , , , ...	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
p50s63nPq9I	[2, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, ...	[welcome, , , , , , or, , , , , , , , , , , ...	[Welcome, , , , , , 4, , , , , , , , , , , , ,...	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...

From `label_diff_targets()`

```
BOTH_AGREE = 0
BOTH_DIFFER = 2
AUTOGEN_INSERT = 1
MANUAL_INSERT = -1
```

Dataset: Jeremy Howard example

```
In [41]: transcripts.loc['_QUEXsHfsA0']['diffs'][40:70]

Out[41]: ['it',
          'to',
          'you',
          'live',
          'from',
          'day',
          'number',
          'one',
          'of',
          'a',
          'complete',
          {'+': ['shutdown.', 'Oh,'], '-': ['shutdown', 'or']},
          'not',
          {'+': ['a', 'complete', 'shutdown,', 'but', 'nearly', 'a'], '-': []},
          'complete',
          'shutdown',
          {'+': ['of'], '-': ['but', 'nearly', 'complete', 'shutdown', 'in']},
          'San',
          {'+': ['Francisco.', "We're"], '-': ['Francisco', "we're"]},
          'going',
          'to',
          'be',
          'recording',
          'it',
          'over',
          'the',
          {'+': [], '-': ['defect']},
          'next',
          'two',
          'months']
```

https://github.com/2dot71mily/youtube_captions_corrections

Dataset: Add your fav channels

Adding_a_youtube_channel_to_yt_caption_corrections_dataset

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Editing

```
[5] %%capture
    !git clone https://github.com/2dot71mily/youtube_captions_corrections.git
    !pip install youtube_transcript_api
    !pip install google-api-python-client
```

```
[6] !python youtube_captions_corrections/src/prepare_data.py
```

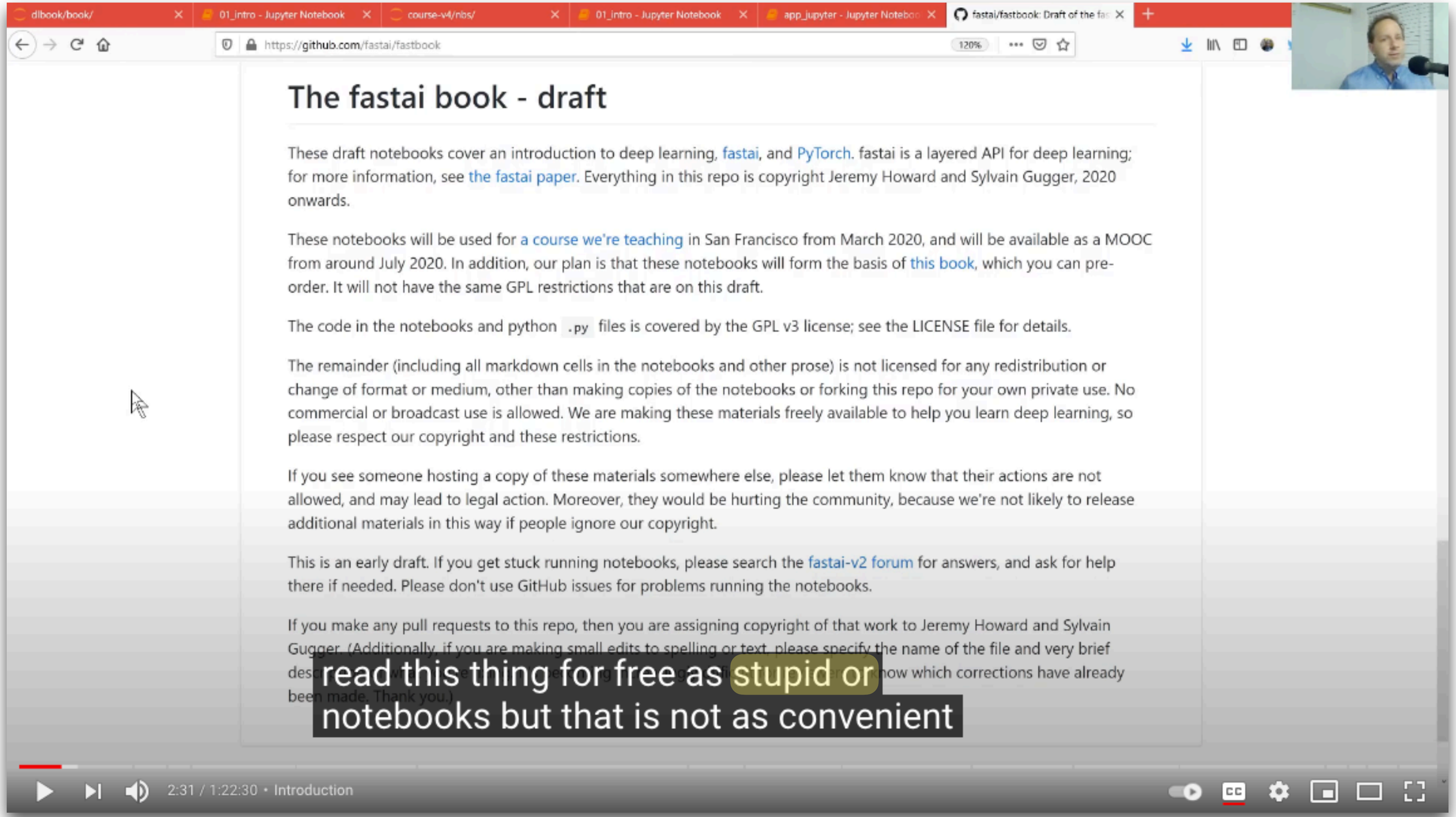
```
requesting data for /Users/emcmilin/youtube_captions_corrections/temp/data/transcripts/en/raw_transcripts/Jeremy_Howard.json
requesting data for /Users/emcmilin/youtube_captions_corrections/temp/data/channels/Jeremy_Howard.json

        If below requested channel is correct:
        https://www.youtube.com/channel/UCX7Y2qWriXpqocG97SFW2OQ
        please hit [Enter]

        Else [ctrl+c] script and check for other `channel_id` options in:
        /Users/emcmilin/youtube_captions_corrections/temp/data/channels/Jeremy_Howard.json

requesting data for /Users/emcmilin/youtube_captions_corrections/temp/data/playlists/Jeremy_Howard.json
requesting data for /Users/emcmilin/youtube_captions_corrections/temp/data/videos/Jeremy_Howard.json
requesting data for /Users/emcmilin/youtube_captions_corrections/temp/data/transcripts/en/raw_transcripts/Jeremy_Howard.json
requesting data for /Users/emcmilin/youtube_captions_corrections/temp/data/transcripts/en/labeled_transcripts/Jeremy_Howard.json
```

config.py

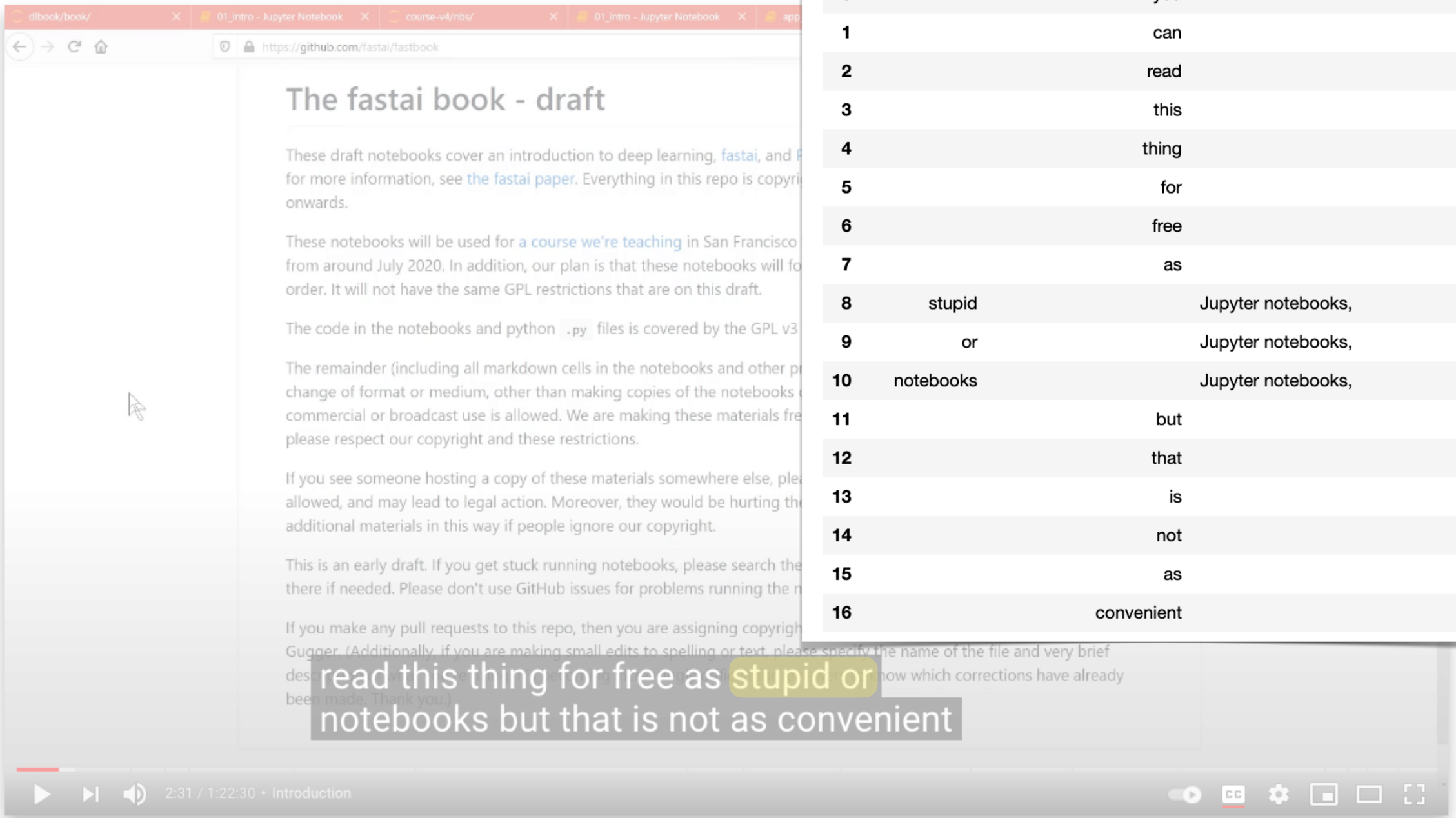
```
1  from pathlib import PurePath
2
3  ### Prefs ###
4  DEVELOPER_KEY = "[redacted]"
5  LANGUAGE = "en"
6  CHANNEL_NAME = "Jeremy Howard"
7  GET_CHANNEL_IDS_ONLY = False
8  GET_PLAYLIST_IDS_ONLY = False
9  GET_VIDEO_IDS_ONLY = False
10
11 SAVE_INTERVAL = 100
12 TRANSCRIPT_SAVE_INTERVAL = SAVE_INTERVAL
13
14 PRINT_TRANSCRIPT_API_ERR = False
15 USE_ONLY_POSTPROC_LABELS = True
16 USE_VIDEO_ID_AS_IDX = False
17
18 SPLIT_FILE_N_LINES = 3000
19
20 ### Labels ###
```

https://colab.research.google.com/drive/1pRNGBEQcQVTLVkXlYfkI0Dlkg3i9jfcz?usp=sharing

Dataset: Revisiting stupid or...?

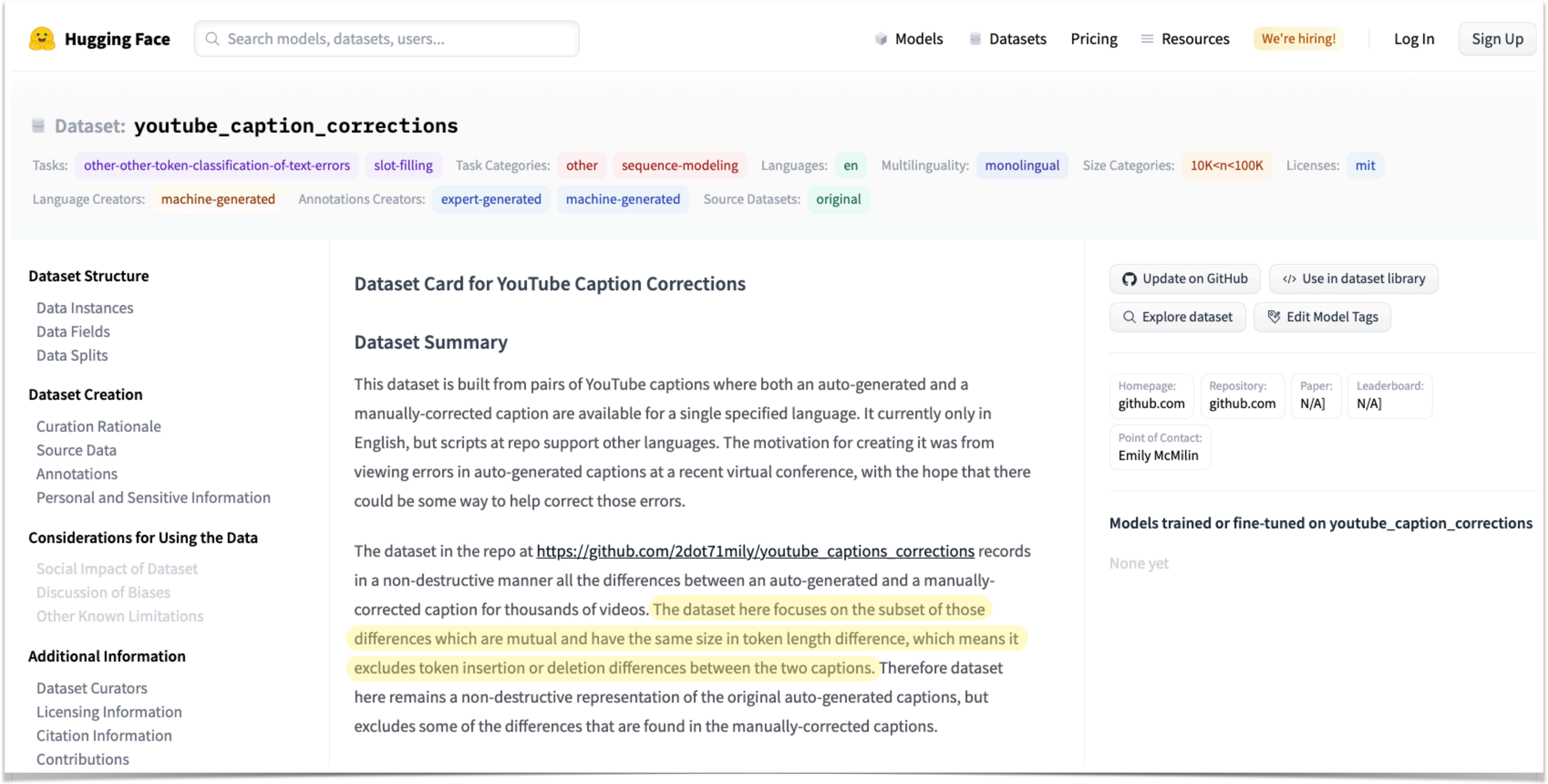


Dataset: Revisiting stupi



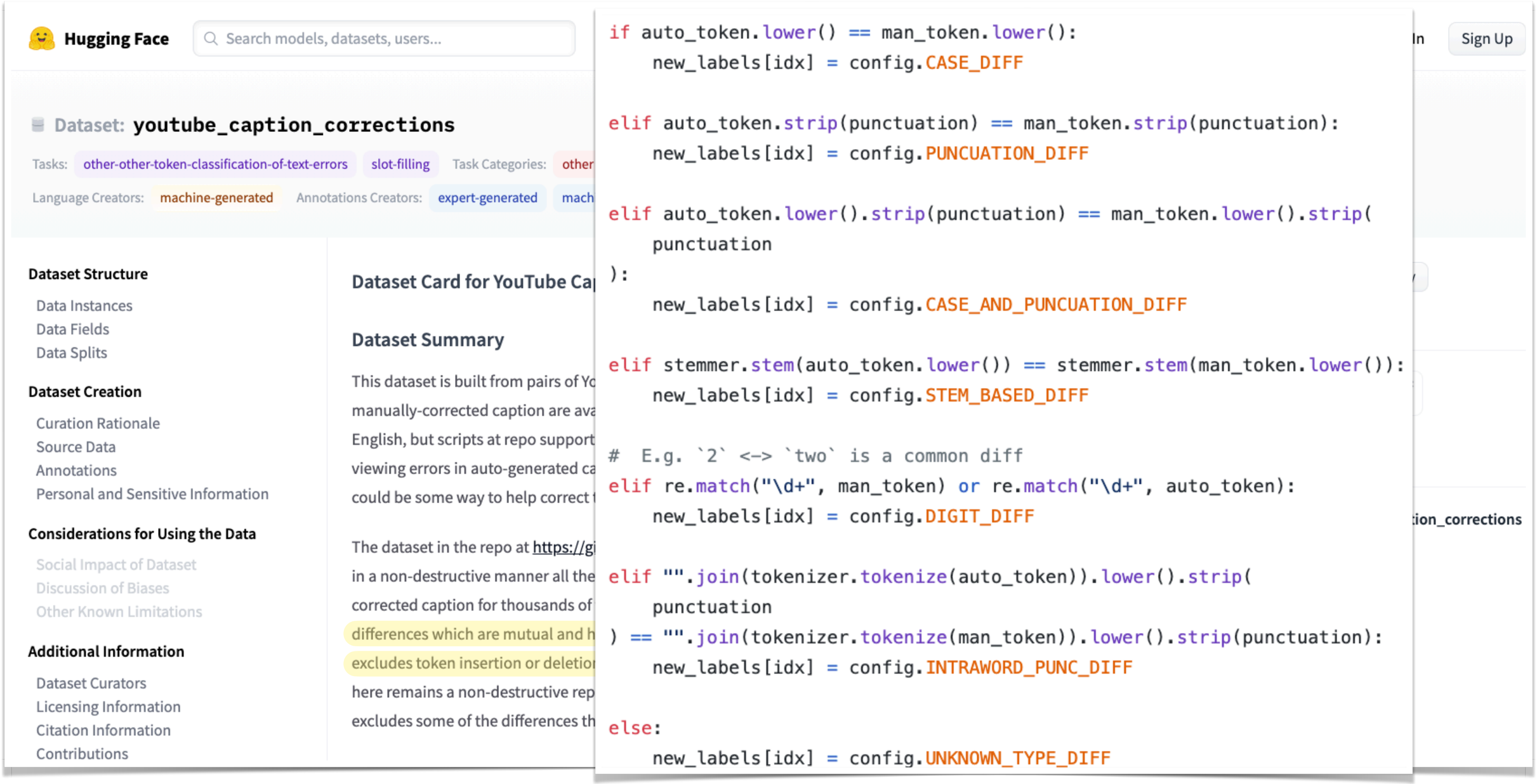
	autogen_seq	common_to_both_seq	manual_seq	manual_addl_rep
0		you		0
1		can		0
2		read		0
3		this		0
4		thing		0
5		for		0
6		free		0
7		as		0
8	stupid		Jupyter notebooks,	2
9	or		Jupyter notebooks,	2
10	notebooks		Jupyter notebooks,	2
11		but		0
12		that		0
13		is		0
14		not		0
15		as		0
16		convenient		0

Dataset: Subset added to HuggingFace



https://huggingface.co/datasets/youtube_caption_corrections

Dataset: Labels added to HuggingFace




```python
if auto_token.lower() == man_token.lower():
    new_labels[idx] = config.CASE_DIFF

elif auto_token.strip(punctuation) == man_token.strip(punctuation):
    new_labels[idx] = config.PUNCUATION_DIFF

elif auto_token.lower().strip(punctuation) == man_token.lower().strip(
    punctuation
):
    new_labels[idx] = config.CASE_AND_PUNCUATION_DIFF

elif stemmer.stem(auto_token.lower()) == stemmer.stem(man_token.lower()):
    new_labels[idx] = config.STEM_BASED_DIFF

#  E.g. `2` <-> `two` is a common diff
elif re.match("\d+", man_token) or re.match("\d+", auto_token):
    new_labels[idx] = config.DIGIT_DIFF

elif "".join(tokenizer.tokenize(auto_token)).lower().strip(
    punctuation
) == "".join(tokenizer.tokenize(man_token)).lower().strip(punctuation):
    new_labels[idx] = config.INTRAWORD_PUNC_DIFF

else:
    new_labels[idx] = config.UNKNOWN_TYPE_DIFF
```

https://huggingface.co/datasets/youtube_caption_corrections

https://github.com/2dot71mily/youtube_captions_corrections/blob/main/src/postprocess_data.py

Model: Token-Classification

Model: Getting dataset from HuggingFace

```
[10] from datasets import load_dataset
```

```
[11] if TESTING:
         yt_dataset = load_dataset("youtube_caption_corrections", split='train[:5%]')
     else:
         yt_dataset = load_dataset("youtube_caption_corrections")['train']
```

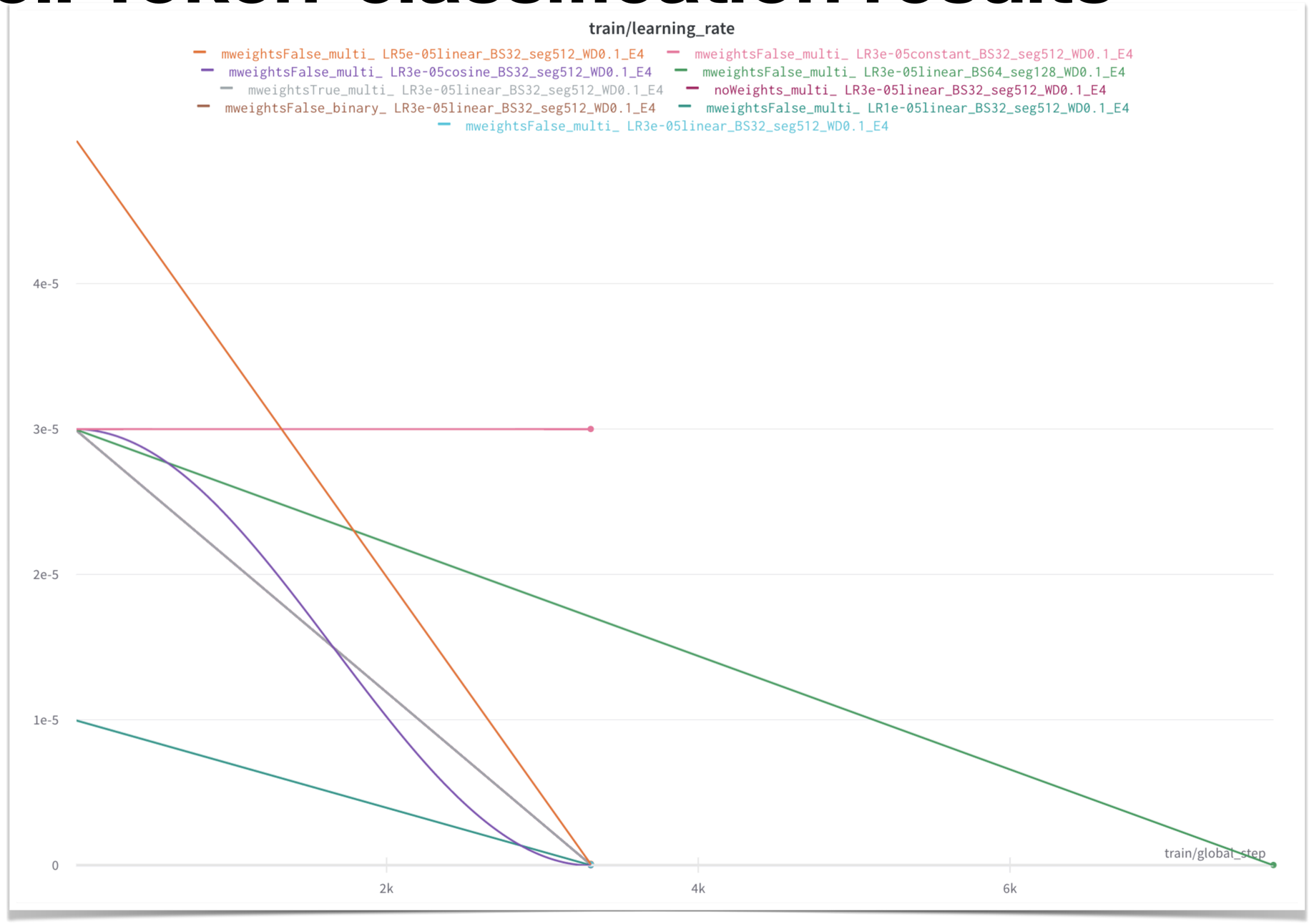
```
Downloading: 4.16k/? [00:06<00:00, 647B/s]

Downloading: 2.86k/? [00:00<00:00, 55.2kB/s]
Using custom data configuration default

Downloading and preparing dataset youtube_caption_corrections/default (download:
Downloading: 59.9M/? [00:11<00:00, 5.16MB/s]

Downloading: 54.3M/? [00:05<00:00, 9.82MB/s]

Downloading: 51.5M/? [00:08<00:00, 6.36MB/s]

Downloading: 56.9M/? [00:01<00:00, 40.4MB/s]

Dataset youtube_caption_corrections downloaded and prepared to /root/.cache/hugg
```

```
[12] yt_dataset
```

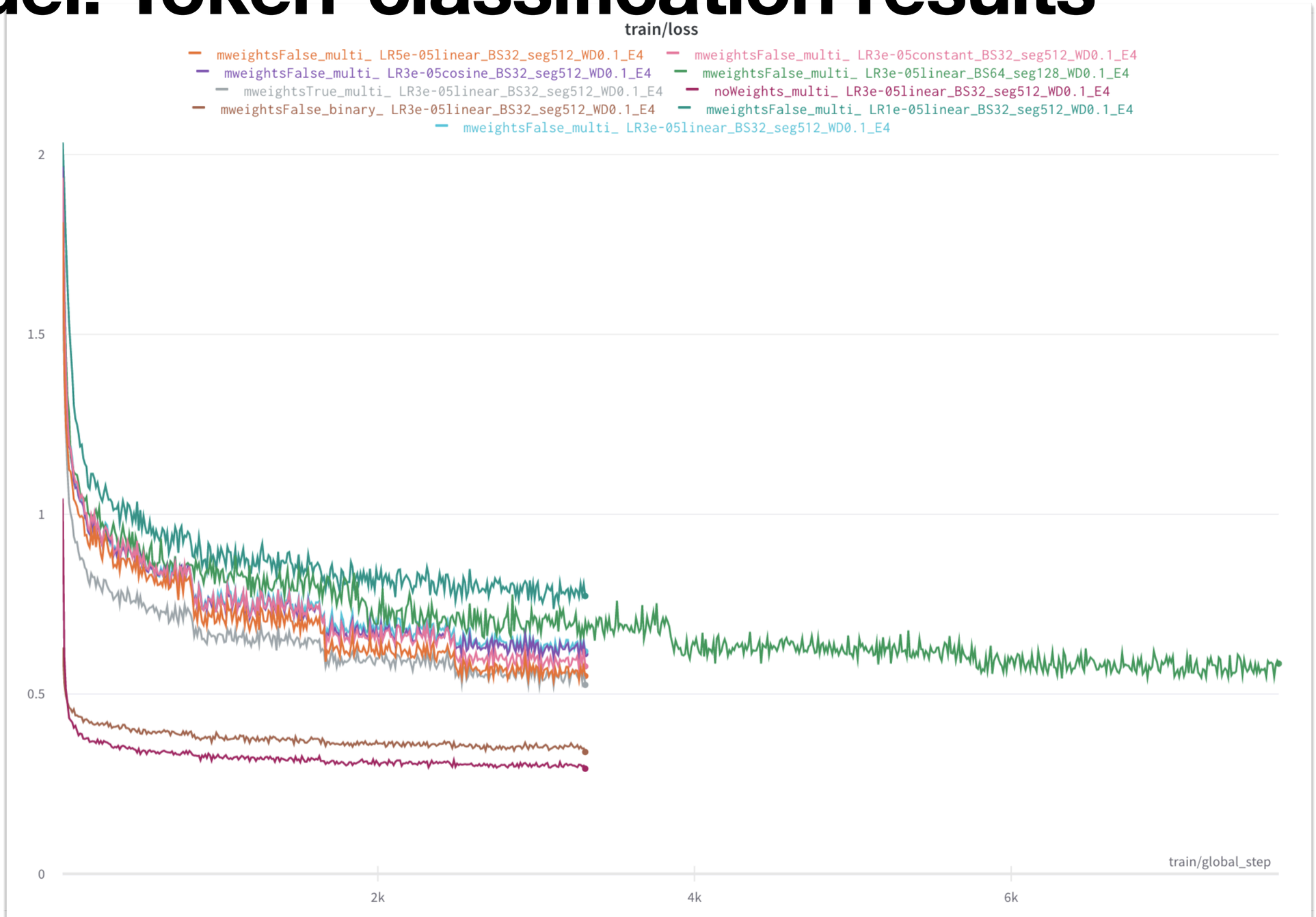
```
Dataset({
    features: ['video_ids', 'default_seq', 'correction_seq', 'diff_type'],
    num_rows: 10769
})
```

https://colab.research.google.com/drive/1olAN3mqunN6FWpgxTgL6W72zdp6Jzbcf?usp=sharing

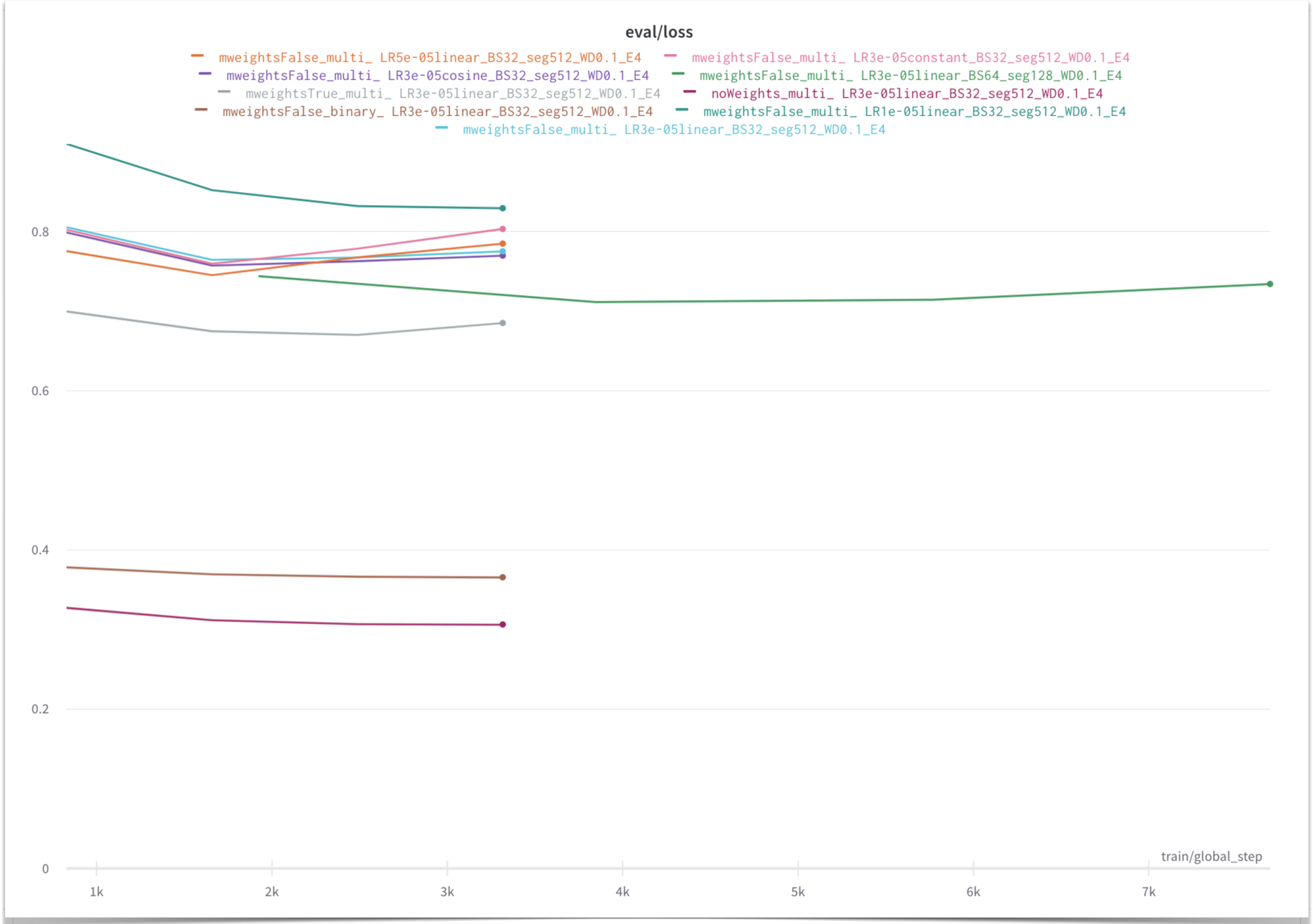
Model: Token-classification results



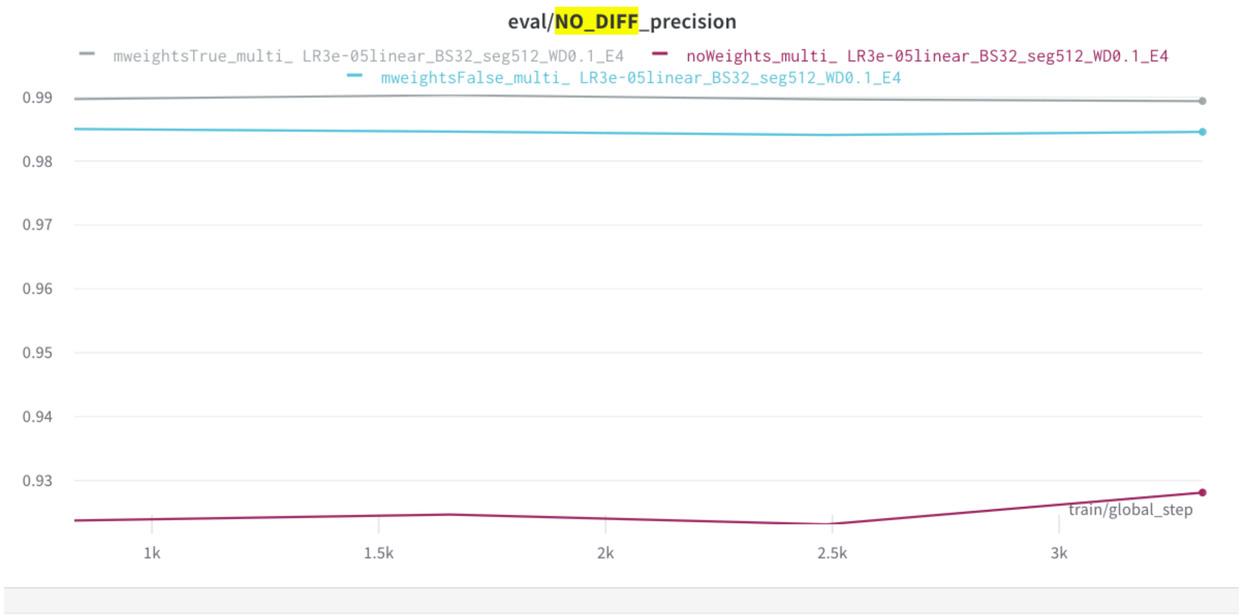
Model: Token-classification results

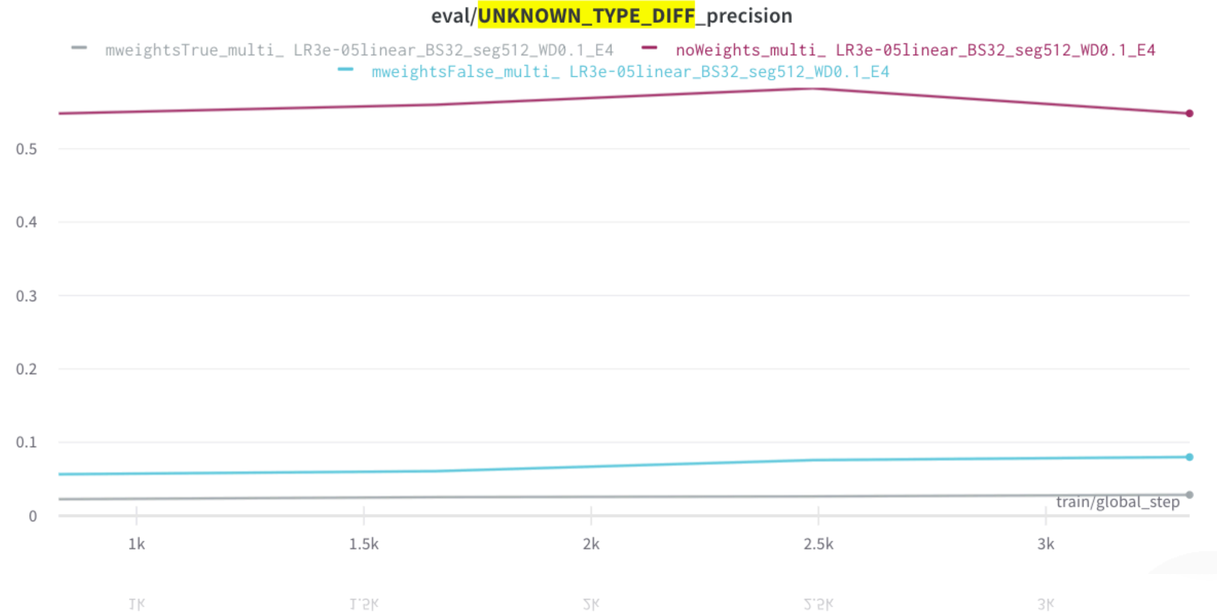


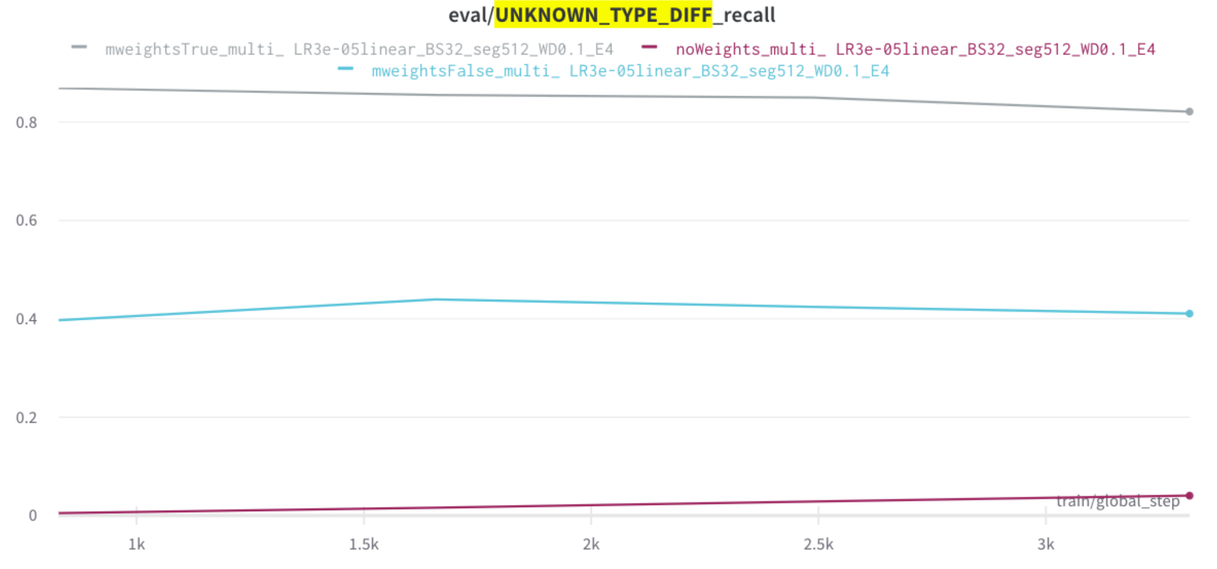
Model: Token-classification results



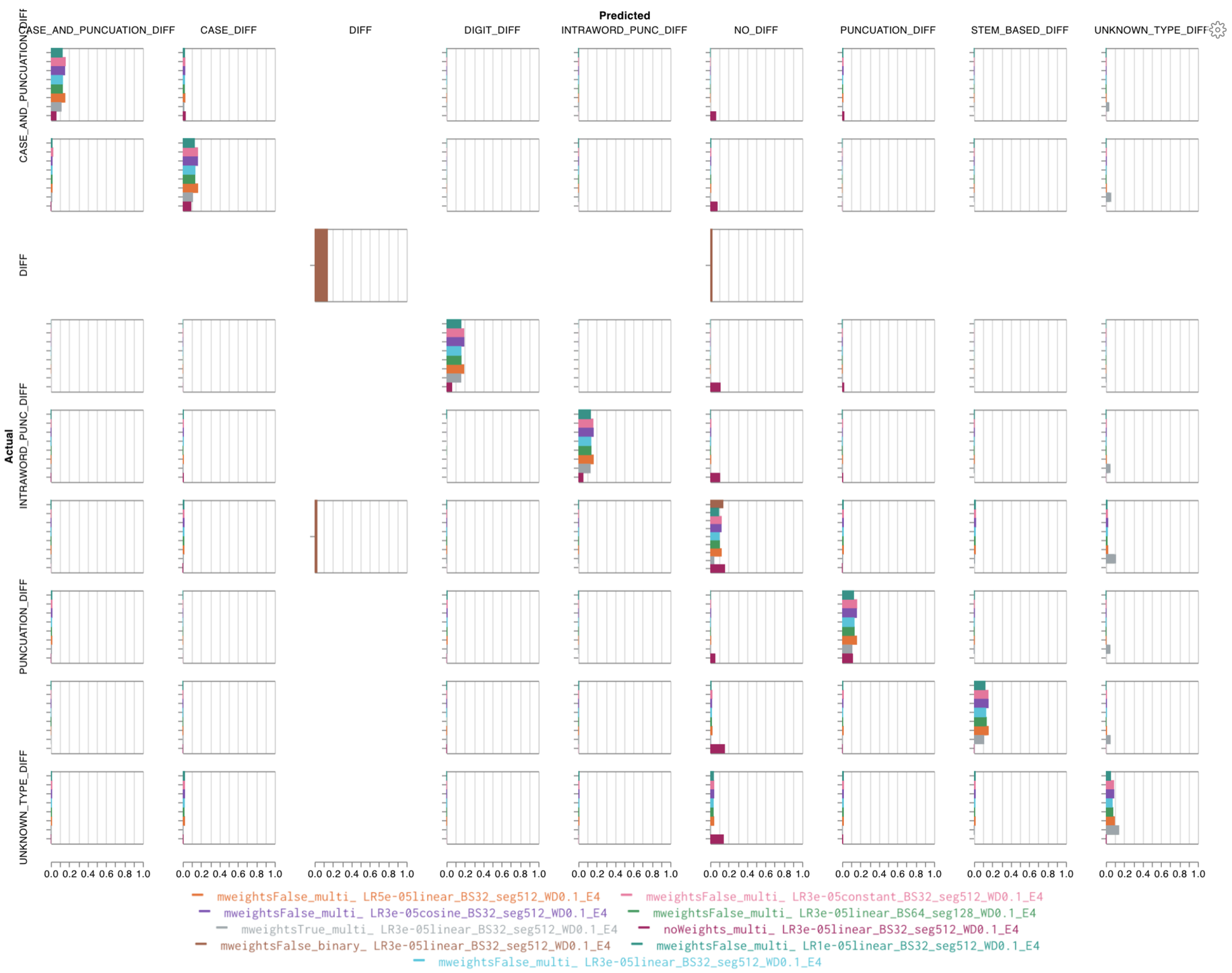
Model: Token-classification results








Predicted
CASE_AND_PUNCUATION_DIFF
CASE_DIFF
DIGIT_DIFF
INTRAWORD_PUNC_DIFF
NO_DIFF
PUNCUATION_DIFF
STEM_BASED_DIFF
UNKNOWN_TYPE_DIFF
Actual
CASE_AND_PUNCUATION_DIFF
DIGIT_DIFF
NO_DIFF
STEM_BASED_DIFF
0.0 0.2 0.4 0.6 0.8 1.0
mweightsTrue_multi_ LR3e-05linear_BS32_seg512_WD0.1_E4
noWeights_multi_ LR3e-05linear_BS32_seg512_WD0.1_E4
mweightsFalse_multi_ LR3e-05linear_BS32_seg512_WD0.1_E4


Predicted
CASE_AND_PUNCUATION_DIFF
CASE_DIFF
DIFF
DIGIT_DIFF
INTRAWORD_PUNC_DIFF
NO_DIFF
PUNCUATION_DIFF
STEM_BASED_DIFF
UNKNOWN_TYPE_DIFF
Actual
0.0 0.2 0.4 0.6 0.8 1.0
mweightsFalse_multi_ LR5e-05linear_BS32_seg512_WD0.1_E4
mweightsFalse_multi_ LR3e-05constant_BS32_seg512_WD0.1_E4
mweightsFalse_multi_ LR3e-05cosine_BS32_seg512_WD0.1_E4
mweightsFalse_multi_ LR3e-05linear_BS64_seg128_WD0.1_E4
mweightsTrue_multi_ LR3e-05linear_BS32_seg512_WD0.1_E4
noWeights_multi_ LR3e-05linear_BS32_seg512_WD0.1_E4
mweightsFalse_binary_ LR3e-05linear_BS32_seg512_WD0.1_E4
mweightsFalse_multi_ LR1e-05linear_BS32_seg512_WD0.1_E4
mweightsFalse_multi_ LR3e-05linear_BS32_seg512_WD0.1_E4

Next steps

Dataset
- Increase dataset size
- Include insertion/deletion errors in HuggingFace dataset

Token classification model:
- Train different models for different types of errors?
- Improve loss function to weigh important error types more heavily

Masked Language Model:
- Train model with MLM objective:
  - Mask out words labeled incorrect
- Attempt an end to end model:
  - Mask out words predicted as incorrect