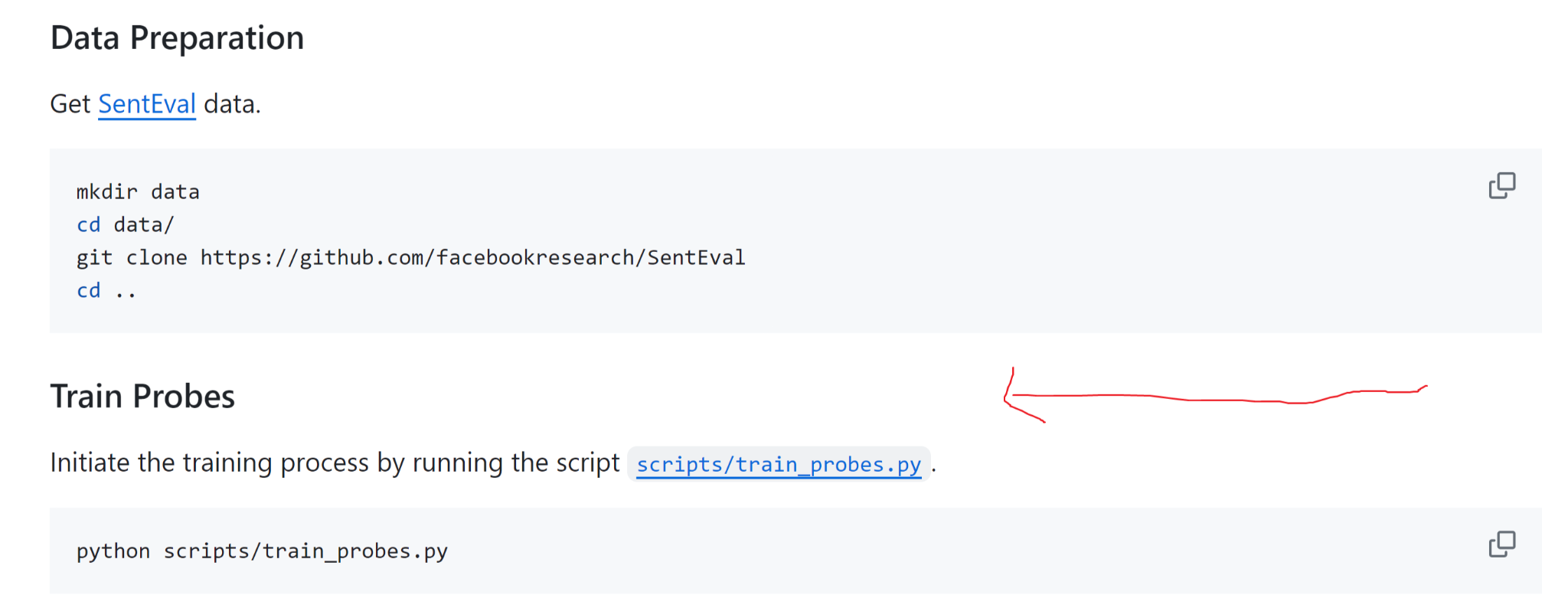
Link to poster:<https://docs.google.com/presentation/d/1-6uei8ANbfR7Ek3GI9fJhjPc2tiJ7g1r/edit?usp=drive_link&ouid=113801578927746711694&rtpof=true&sd=true>

Link to the files to be substituted into the following directories, before you can the ‘conflict\_type’ probing task (This will become clear later): <https://drive.google.com/drive/folders/1Ljl2_1gd-3F9rMDkML1_Is-uacOTz4I0?usp=drive_link>

Tasks left for Kai before the presentation, other than annotation:

Set up as instructed by Niu et al:

<https://github.com/frankniujc/gridloc_probe/blob/main/README.md>

1. Follow the steps till just before you train probes:
2. Substitute the following files in the directory “.../gridloc\_probe/scripts”, you’ll find the modified files in the link above:



1. Add this file to “.../gridloc\_probe/experiments/sent\_eval\_tasks”:

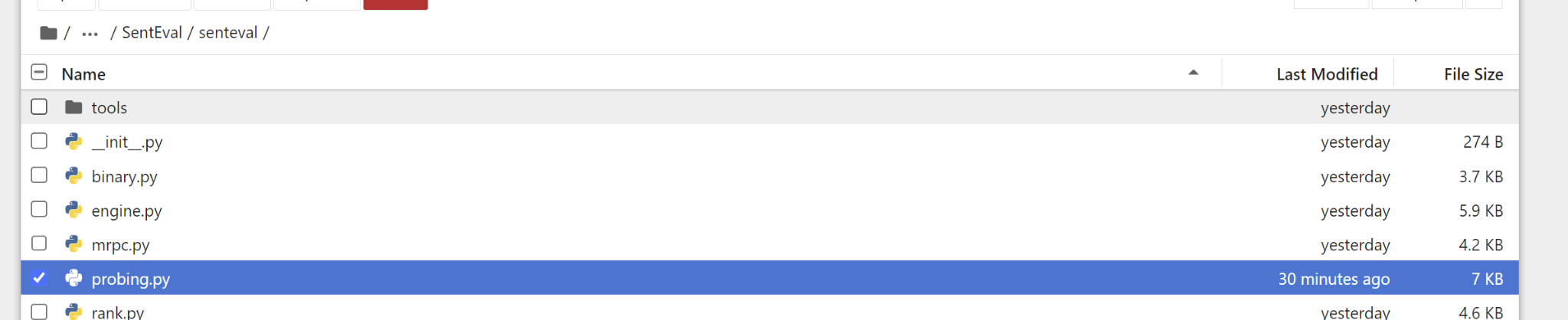


1. Add this file (not provided in the link above) to “ANLP\_final\_project/gridloc\_probe/data/SentEval/data/probing”:



Note you’ll need to convert the annotated data into this txt file. For requirements, refer to this file, “.../gridloc\_probe/data/SentEval/data/probing/README.md”. Note that the separation strings (tr, te, va) in the first column of the txt file, indicate whether the row is training, testing, or validation.

1. Substitute this file in “.../gridloc\_probe/data/SentEval/senteval”:

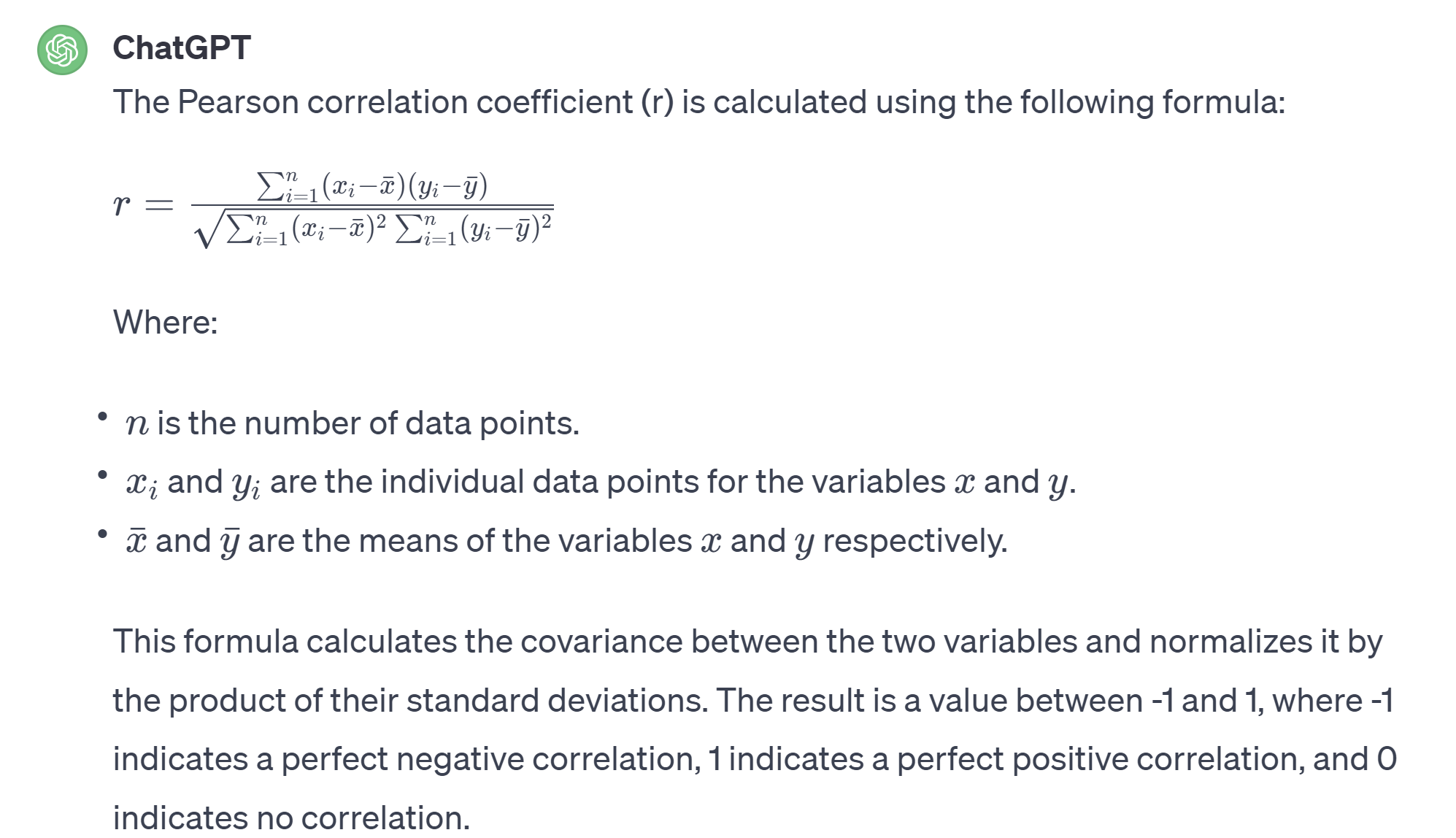


1. Continue with the instruction provided by Niu et al. We can ignore TreeDepth Correlation Analysis.
2. Interpret the results as you see fit, and complete the poster. My suggestion is to pick one or two plots from plot\_sentence, which hopefully shows the punchline lighten up, as well as the set-up with two interpretations for type 0 conflicts.

Statistical hypothesis testing:

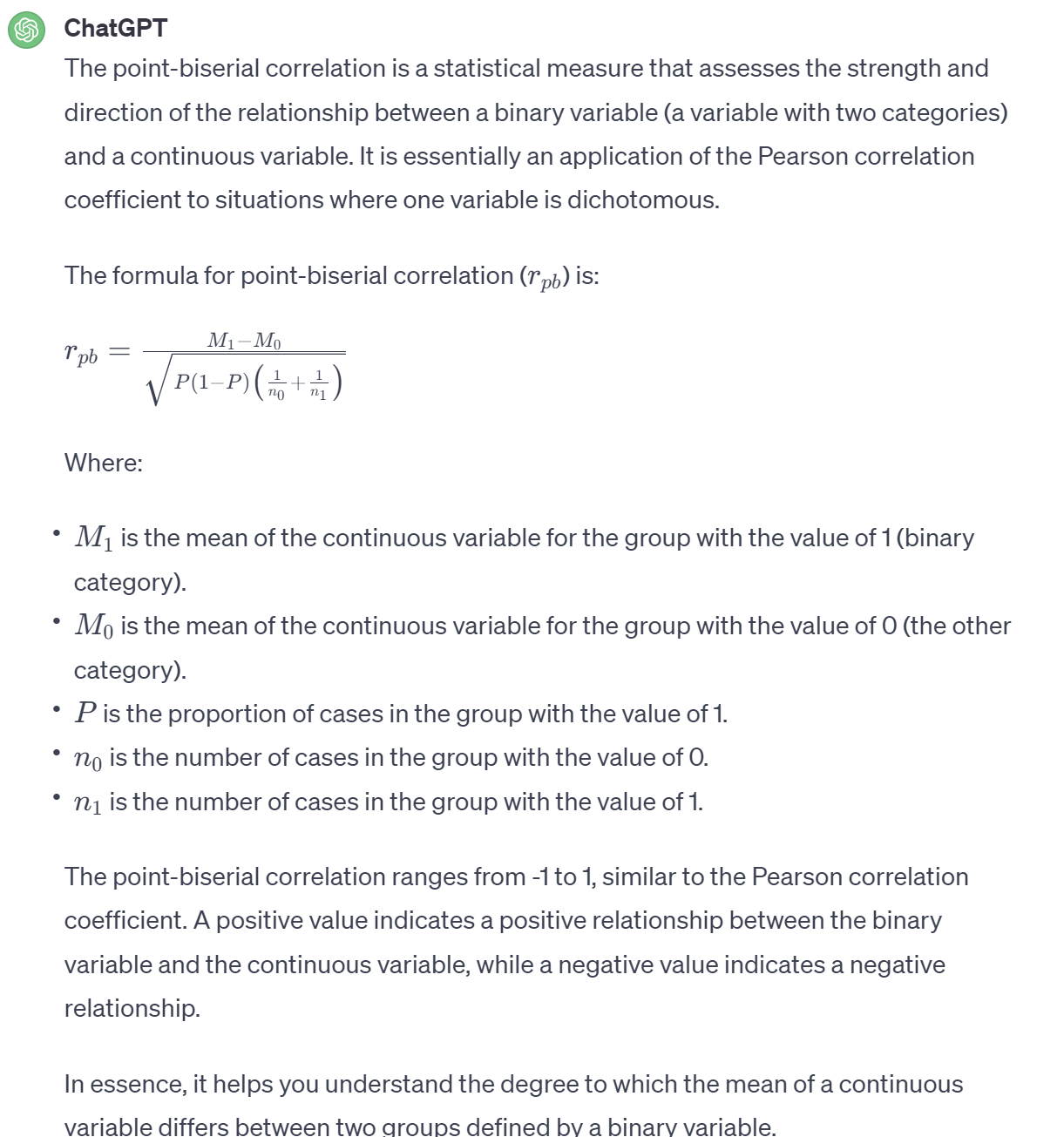
* Train a same sentence over 50-100 times
* Collect token attention weight distribution
* Run a hypothesis test on CLS to joke set up, punchline to joke set up. The mean value should be different
* CLS, Type 1 punchline, Type 0 setup, Type 0 punchline, rest of tokens

Pearson Correlation coefficient.



* For type 1 joke, mark the CLS token, and the punchline tokens as 1, the rest tokens as 0. For type 0 joke, mark the CLS token, the set-up tokens, and the punchline tokens as 1, the rest as 0.

Point-biserial:



* For type 1 joke, mark the CLS token, and the punchline tokens as 1, the rest tokens as 0. For type 0 joke, mark the CLS token, the set-up tokens, and the punchline tokens as 1, the rest as 0.