For our approach we use a Roberta large model pretrained on mnli as our base. For each pair of sentences in the story we perform a binary classification of whether that pair makes sense or not. To determine whether a sentence is a breakpoint, we take a smooth approximation to the maximum function over all sentence pairs that include that sentence and all previous sentences. We then use a threshold to determine if that sentence should be considered a breakpoint or not. If there are multiple breakpoints, the earliest one is returned. For training we use a binary cross entropy loss on the sentence breakpoint predictions (ie the smooth maximum of all pairs… etc stated above). Any sentence pairs before the breakpoint are considered consistent, the breakpoint sentence is inconsistent, and any pair after the breakpoint is ignored because we don’t know.

In terms of performance, if the optimal threshold is chosen for Task 1 we can get 89% validation accuracy on a 10% held-out split. Similarly we can get 73% validation accuracy for Task 2.

To train the model simply run the first 2 cells in the notebook, with the first cell defining the save path and training data path. To perform inference, run the first, third, and fourth cells. The code is currently configured to train all paramters, there are 2 comments with the required code changes to freeze the base model if desired. The optimal threshold is printed during the training stage. All other hyperparameter settings or training decisions are written explicitly in the code