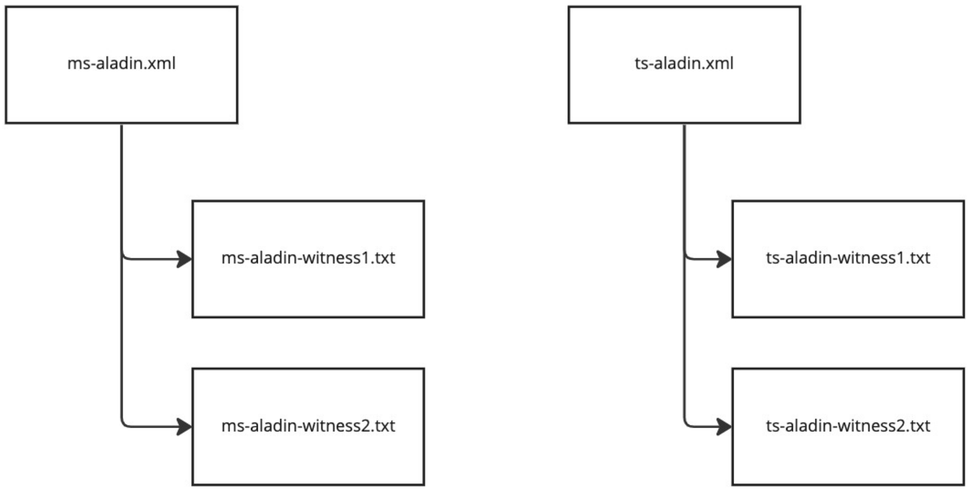
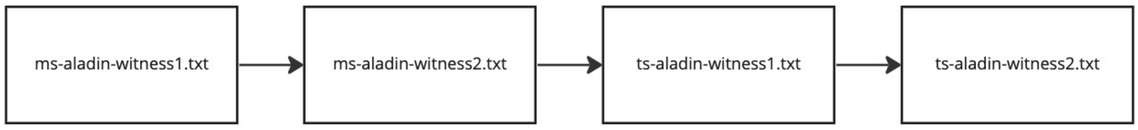
Computation of the witnesses: (witness 1 instant edits, witness 2 all edits)



Chronological sequence of the witnesses:



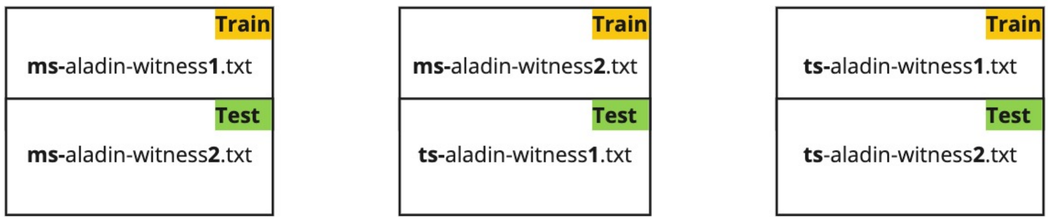
Strategy for using a plagiarism detection algorithm to "predict" alignment of tokens and subsequences of tokens in the witnesses:

1. Select a witness text
2. Use the text from the witness just prior to the selected witness to train the plagiarism detection model.
3. Apply the model to the selected witness to predict the overlap / differences

Motivation: we want to be able to identify overlapping or differing tokens between **each** of the four witnesses. The above strategy is one way to achieve this.

Another strategy could be to train on all other witnesses except for the selected one.

Run plagiarism detection algorithm on data:



Plagiarism detection results analysis:

The main idea behind the use of plagiarism detection is that it can identify phrases and tokens across witnesses that are very similar. These overlapping or similar phrases would be the text that is common to both witnesses. Conversely, by identifying the similar phrases the remaining phrases would be indications of changes or differences among the witnesses. The nice thing about the plagiarism detection algorithm is that it gives a score from 0.0 - 1.0 indicating the similarity of the token meaning (context) across witnesses. A score of 1.0 indicates that the token shares exactly the same context across the two input witnesses. A score of 0.0 indicates that the contexts of this token in the two witnesses are completely unrelated. **The questions we have are:**

1. **What is the accuracy and precision of the algorithm with regards to identifying the same semantic context across two witnesses?**
2. **If the accuracy and precision of the algorithm is satisfactory for identifying the same semantic context across two witnesses, then what is the threshold moving from 1.0 down to 0.0 at which we arrive at completely different semantic contexts, where we do not want to "align" tokens across these contexts.**