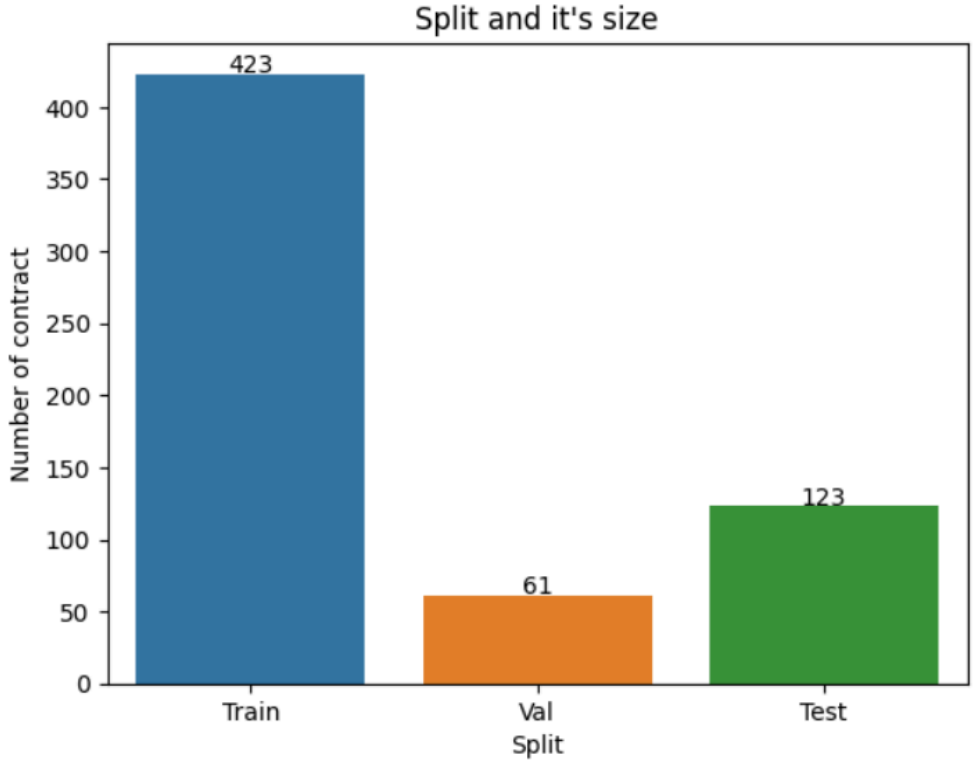
❖ EDA - Exploratory data analysis

➢ There are total 17 hypothesis in train, dev and test split.



➢ There is not a single entry with a NULL value in all three splits so the data is clean.

➢ Most of the contract is length 5000 to 15000 (in character). The avg contract length is 11,050 (in character) in train data

➢ Also for dev and test data, the average contract length is around 11,200 (in character).

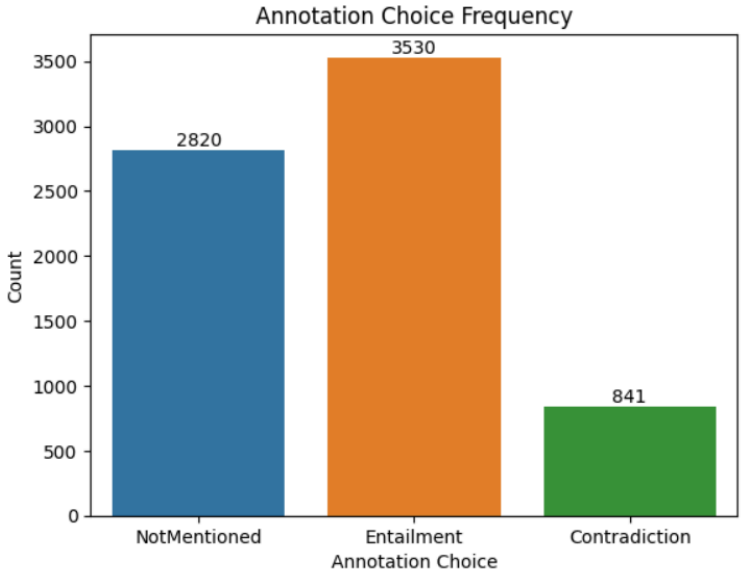
➢ Most of the span length is less than 400. Same in dev and test data.

➢ The annotation count per contract is 17 for all contracts in train, dev , and test.

➢ There are only 3 types of contract:- 'search-pdf', 'sec-text', 'sec-html'.

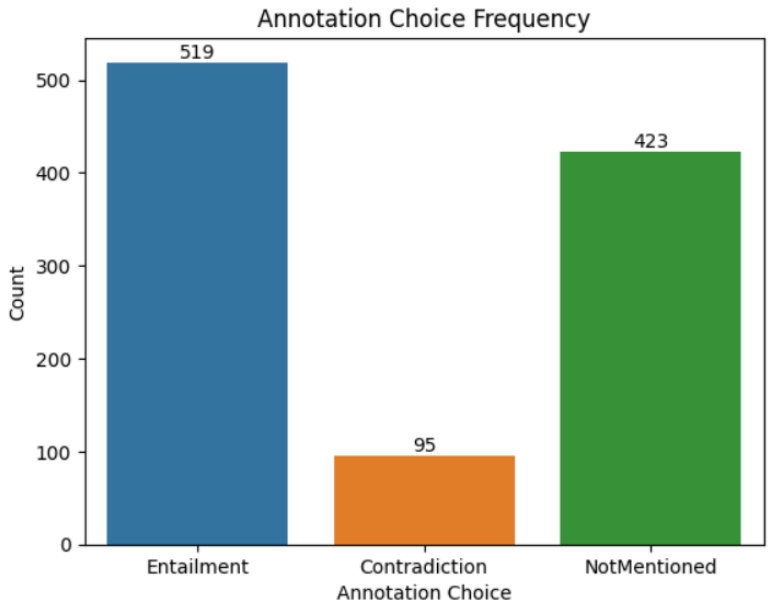
➢ ‘Search-pdf’ is the most frequent

➢ **Train Dataset Annotation distribution**



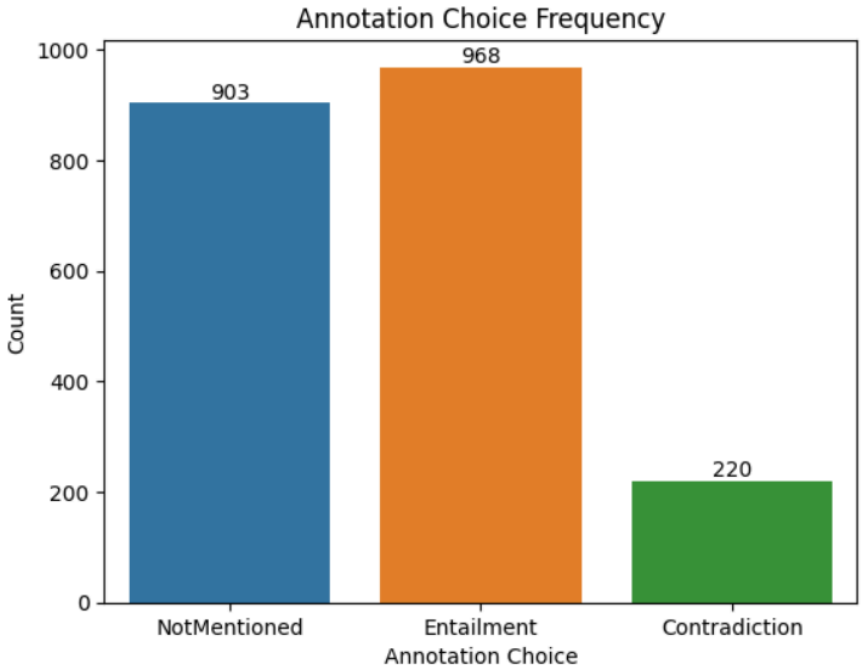
- The train data is imbalanced because the ratio of classes Entailment: Contraction: Not Mentioned is almost 4:1:3.

➢ **Dev Dataset Annotation distribution**



- The dev data is imbalanced because the ratio of classes Entailment: Contraction: Not Mentioned is almost 5:1:4.

➢ **Test Dataset Annotation distribution**



- The test data is imbalanced because the ratio of classes Entailment: Contraction: Not Mentioned is almost 4:1:4.

- Loss scaling or undersampling of the oversampled class or oversampling of the undersampled class is required because there is a remarkable class imbalance.