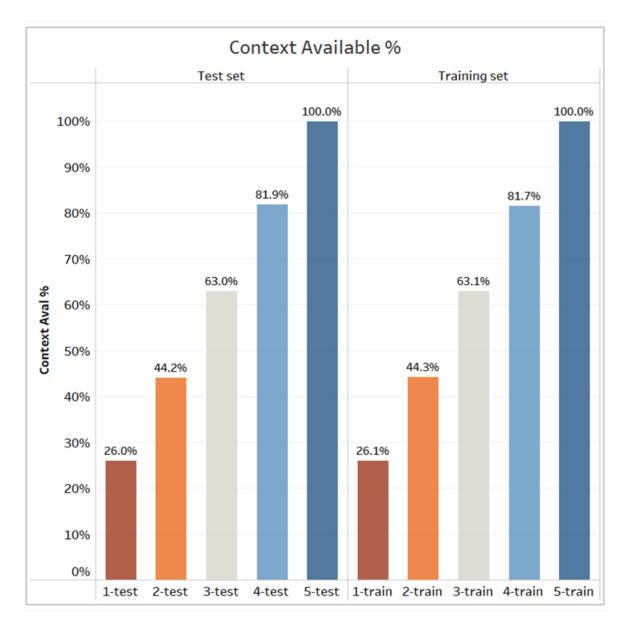
BERT for Question Answering in Biomedical Domain



Context availability in terms of context length in the training and testing set

Figure : Training and Testing set context availability percentage

Answerable examples with varying context length (Original Results)

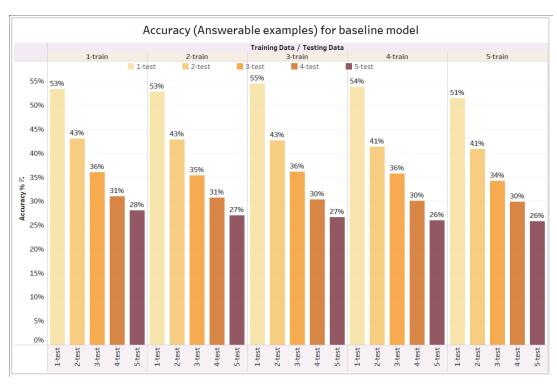


Figure : Accuracy of the baseline model with varying context length

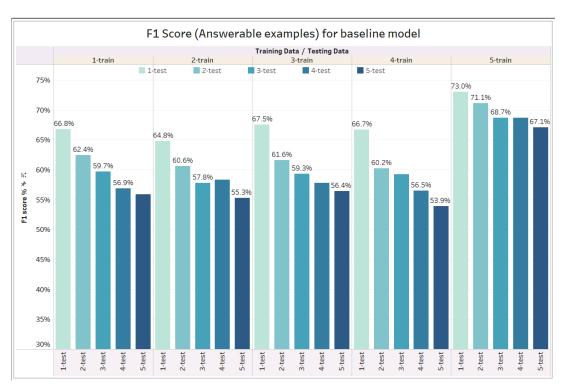
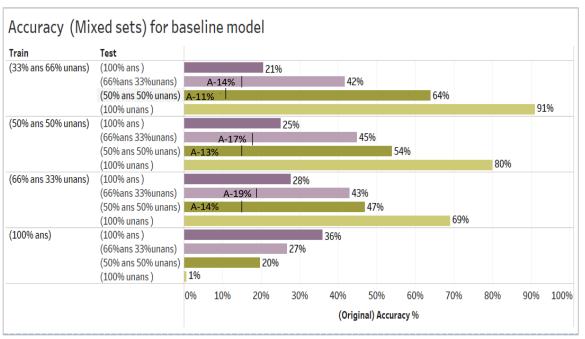


Figure : F1 score of the baseline model with varying context length

Mixed examples (Original Results)



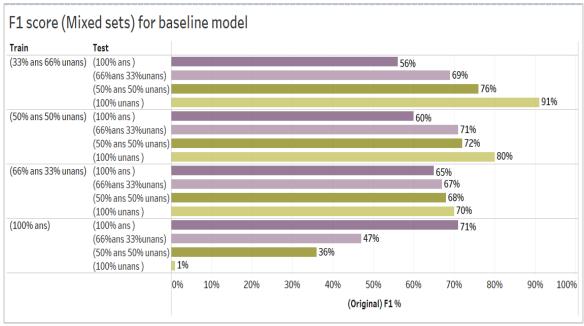


Figure: Accuracy with mixed examples

Figure : F1 score with mixed examples

Span Selection technique

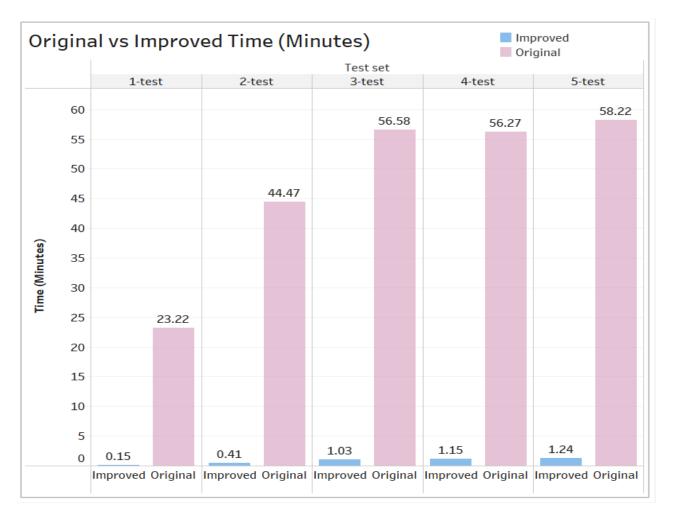


Figure : Comparison of Original and Improved Time (Examples with Varying context length)

Examples with varying context length(Original and improved)

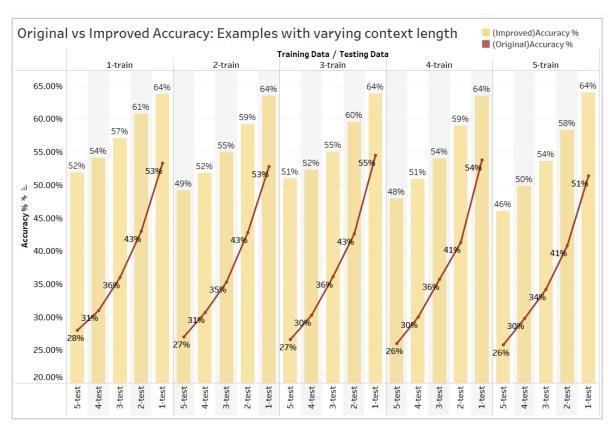


Figure : Comparison of Original and Improved Accuracy (Examples with Varying context length)

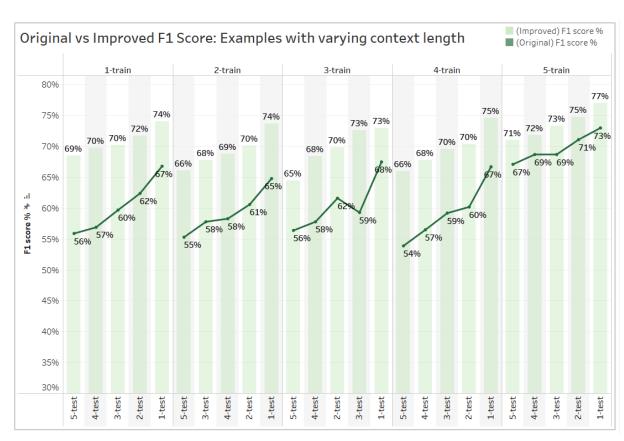
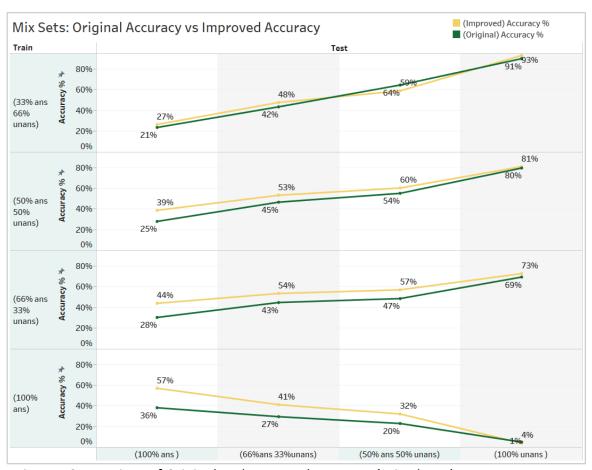


Figure : Comparison of Original and Improved F1 score (Examples with Varying context length

Mixed set (Original and improved)



Mix Sets: Original F1 Score vs Improved F1 Score (Improved) F1 % (Original) F1 % Train Test 100% 76% 91% 76% 69% (33% ans 66% unans) 20% 100% 81% 75% 80% 71% 72% (50% ans 50% unans) 100% 76% 76% 74% 80% 70% 68% 60% 67% (66% ans 33% unans) 20% 100% 80% 60% (100% 40% ans) 20% (100% ans) (66% ans 33% unans) (50% ans 50% unans) (100% unans)

Figure: Comparison of Original and Improved Accuracy (Mixed sets)

Figure: Comparison of Original and Improved F1 score (Mixed sets)