Table 1: Table shows the result of Experiment 4. The colors indicate the model configurations that were examined. The results are based on the average viability each counterfactual a model produces across all factuals that were tested.

	Unnamed: 0	experiment	rank	likelihoo
Short Name				
CBGW-CBG-CBGW-IM	249.500000	1.000000	25.500000	0.39380
EGW-ES-EGW-SBI-ES-OPC-SBM-FSR-IM	1249.500000	1.000000	25.500000	0.49671
EGW-ES-EGW-SBI-ES-OPC-SBM-HR-IM	1749.500000	1.000000	25.500000	0.49799
EGW-ES-EGW-SBI-ES-OPC-SBM-RR-IM	2249.500000	1.000000	25.500000	0.49933
RGW-RG-RGW-IM	749.500000	1.000000	25.500000	0.31499

0.1 Determine the best Generator Algorithm

0.1.1 Experimental Setup

Knowing the , we compare the evolutionary algorithm with other algorithms.

In this comparison, we employ the other models mentioned in **??**. Namely, the *Case-Based Generator* and the *Random Generator*.

For the evolutionary algorithm, we choose the model-configuration from **??**, the rate-configuration determined in **??** and the termination point from **??**. Furthermore, we randomly sample **[20]** factuals from the test set and use the same factuals for every generator. We ensure, that the outcomes are evenly divided. The remaining procedure follows the established procedure of previous experiments.

0.1.2 Results

the results shown in Figure 1 show that the evolutionary algorithm **[model-specifier]** returns better results on average. The worst model is the random generated model.

Table 1 shows the detailed results.

0.1.3 Discussion

These results show that the model **[SBI-ES-OPC-FSR]** is clearly superior to the other models. This result is unsurprising, as the baselines do not actively search for an optimal solution. However, knowing these results, a couple of questions remain. Namely, whether the results remain consistent for longer sequences and for other datasets? **Furthermore, how does**

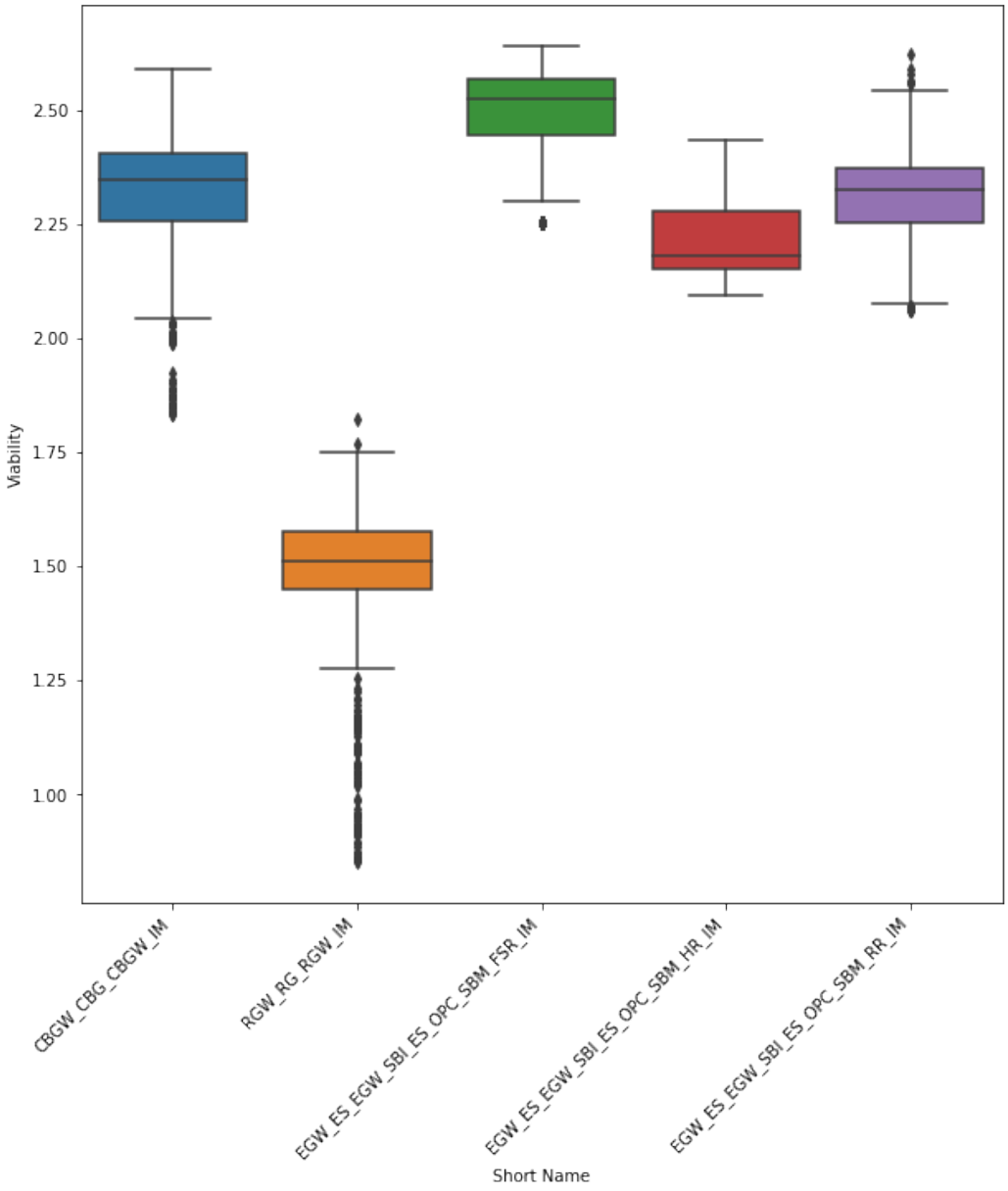

Figure 1: This figure shows boxplots of the viability of each models' generated counterfactual.

this procedure compare to other methods in the literature? The remaining experiments will address these questions.