As mentioned in **??**, counterfactual generation is notorious for lacking a standardised evaluation procedure. Nonetheless, we attempt to address our research questions with the following experiments.

Experiment 1: Model Selection

Before comparing models, we reduce the number of possible models that *can* be compared. In terms of operators, we introduced three initiators, three selectors, three crossers, two mutators and three recombiners. Hence, comparing all possible evolutionary operator combinations requires examining a total of 54 different models. Furthermore, each model has hyperparameters we have to define, too. Therefore, the first set of experiments is dedicated to choosing among a subset of operator combinations and selecting appropriate hyperparameters.

First, we compute all possible configurations without changing any hyperparameter. We refer to each unique operator combination as a model configuration to avoid confusion. For instance, one model configuration would consist of *a Sampling-Based-Initiator, an Elitism-Selector, a One-Point-Crosser, Sampling-Based-Mutator and a Fittest-Survivor-Recombiner*. For the sake of brevity, we refer to a specific model configuration in terms of its abbreviated operators. For instance, the earlier example is denoted as *SBI-ES-OPC-SBM-FSR*.

Afterwards, we explore the hyperparameters of the model. We start with the termination point. Hence, we want to examine the effects of the iterative cycles that each evolutionary algorithm will run. The goal is to find a stopping criterion which yields reasonably good counterfactuals while reducing the computation time. We will only consider the number of iterative cycles as a stopping criterion. We refer to each different criterion as a termination point. Hence, a termination point at 5 means the algorithm will not proceed to optimise its results further after reaching the fifth iteration. We can choose the termination point by inspecting how the average population viability evolves across each cycle. We keep every other experimental setting as established beforehand.

For determining the mutation rate for every mutation type, we choose the best evolutionary algorithm and run the configuration with six rates from 0 to 0.5 in steps of 0.1. We omit everything beyond 0.5 to preserve information about the parent. For instance, if we use a change rate of 0.9, we mutate 90% of the genes the child inherited. This would defeat the purpose of evolving better counterfactuals through breeding. We use the termination point established in the prior experiment. We keep every other experimental setting as set beforehand.

After executing all preliminary experiments, we choose the evolutionary generators and compare them with all baseline models in all subsequent experiments.

Experiment 2: Comparing with Baseline Generators

In this experiment, we assess the viability of all the chosen evolutionary and baseline generators. For this purpose, we sample 10 factuals and use the models to generate 50 counterfactuals. We determine the median viability across the counterfactuals. With this experiment, we show that a model that optimises counterfactual quality criteria produces better results than models that do not. Hence, we expect the evolutionary algorithm to perform best, as it can directly optimise multiple viability criteria. We move on with the best-performing models.
Under *RQ1-H1* and *RQ1-H2* we expect the evolutionary algorithms to outperform the baselines when it comes to viability.

Experiment 3: Comparing with alternative Literature

The model comparison is not enough to establish the validity of our solution, as we defined the viability measure ourselves. Therefore, we also assess each model based on the evaluation criteria of an alternative work. More precisely, we quantify the viability of our models using the metrics employed by Hsieh, Moreira, and Ouyang. Hence, we measure the sparsity by computing the average Levenshstein difference and proximity using the L2-Norm. Furthermore, we compute the average intra-list-diversity and plausibility

Similar to Hsieh, Moreira, and Ouyang, we focus on the *activities* that are generated by each model and its accompanying *resource* event-attribute. For diversity and plausibility, we remain close to the original evaluation protocol by Hsieh, Moreira, and Ouyang as we also treat each counterfactual trace sequence as a symbol. Hence, a sequence *ABC* is treated as a completely different symbol than *ABCD*.

The goal is to show that models, which optimise viability criteria, perform better, even if viability is assessed differently, as stated in *RQ2-H1* of our research question (**??**).

Experiment 4: Qualitative Assessment

For the last assessment, we follow Hsieh, Moreira, and Ouyang's procedure of assessing the models qualitatively. We use the dataset as the authors do. However, as we focus on outcome prediction, we attempt to answer one of two questions:

1. *what would I have had to change to prevent the cancellation/rejection of the loan application process*

2. *what would I have had to change to cause a cancelled/rejected loan application process*

The goal is to show that the results are viable despite not having a standardised protocol to measure their viability.