Hsieh et al. follow a very similar pattern of assessing the quality of their counterfactuals. The authors also focus on the aspects similarity, sparcity, feasibility and likelihood improvement. However, they incorporate and operationalize them differently. Their approach is mostly apparent in their loss function.

[Maybe write the losses in title-case]

Similarity: Similar to our approach, the authors use a distance function and optimize it using gradient descent. They evaluate the quality of their counterfactuals using the same function[1]. However, we use a modified Damerau-Levenshstein distance algorithm to also incorporate structural differences like overlapping or transposed events.

Sparcity: The authors do not optimize towards sparcity, but assess it during their evaluation.

Feasibility: This quality criterion is is embodied by two loss functions: category Loss and scenario Loss. The category loss ensures that categorical variables remain categorical after generation. The scenario loss adds emphasis on only generating counterfactuals that are in the event log. Unlike our probablistic interpretation, they treat the existence of feasible counterfactuals as a binary criterion[2].

Likelihood: Similar to the authors' scenario loss, they treat the improvement of a class as a binary state. Either the counterfactual changes the model's prediction to the desired class or it does not.

The details of each criterion's operationalisation, are explained in [1]. By assessing their interpretation of quality criterions, we see the clear distinction between our approach and the approach of Hsieh et al.

First, their viability measure decisively discourages the generation of counterfactuals that are not present in the dataset. In constast, our approach treats this aspect as a soft constraint.

Second, while our approach also acknowledges general improvements in likelihoods, DiCE4EL treats all counterfactuals that do not lead to better desired as dentrimental solutions. However, one can argue that improving the likelihood of a desired outcome just slightly is already beneficial.

Third, [1] do not optimize sparcity, while we include it within our framework. One can argue that similarity automatically incorporates aspects of sparcity, but we disagree with this notion. We can see this by employing a

[1] They call it proximity during evaluation

[2] They call it plausibility during evaluation

simple example: **[This example can be improved, either by referencing a former example or explaining A, B, C and the event attributes at the begiining.]** Let factual A have features signifying the biological sex (binary), the income (normalized) and the age (normalized) $\begin{pmatrix}1\\1\\1\end{pmatrix}$ as event attributes. Let counterfactual B have the same event attributes with $\begin{pmatrix}0\\1\\1\end{pmatrix}$. Let's assume the distance measure uses the L1-norm. Then, a counterfactual C with event attributes $\begin{pmatrix}1\\0.5\\0.5\end{pmatrix}$, would have the same distance to factual A than B has. However, C requires the change of two event attributes, while B only requires 1 change. In this scenario, B is more preferable than C, despite their distances to A.

The last difference stems from the fact that Hsieh et al. do not include structural sequence characteristics in their similarity measure. A sequence **XXZXX** might be more similar to **XXXZX**, than **XXXXZ**. The former requires only a transposition, while the latter requires two changes. Both have two positions that are not correct.