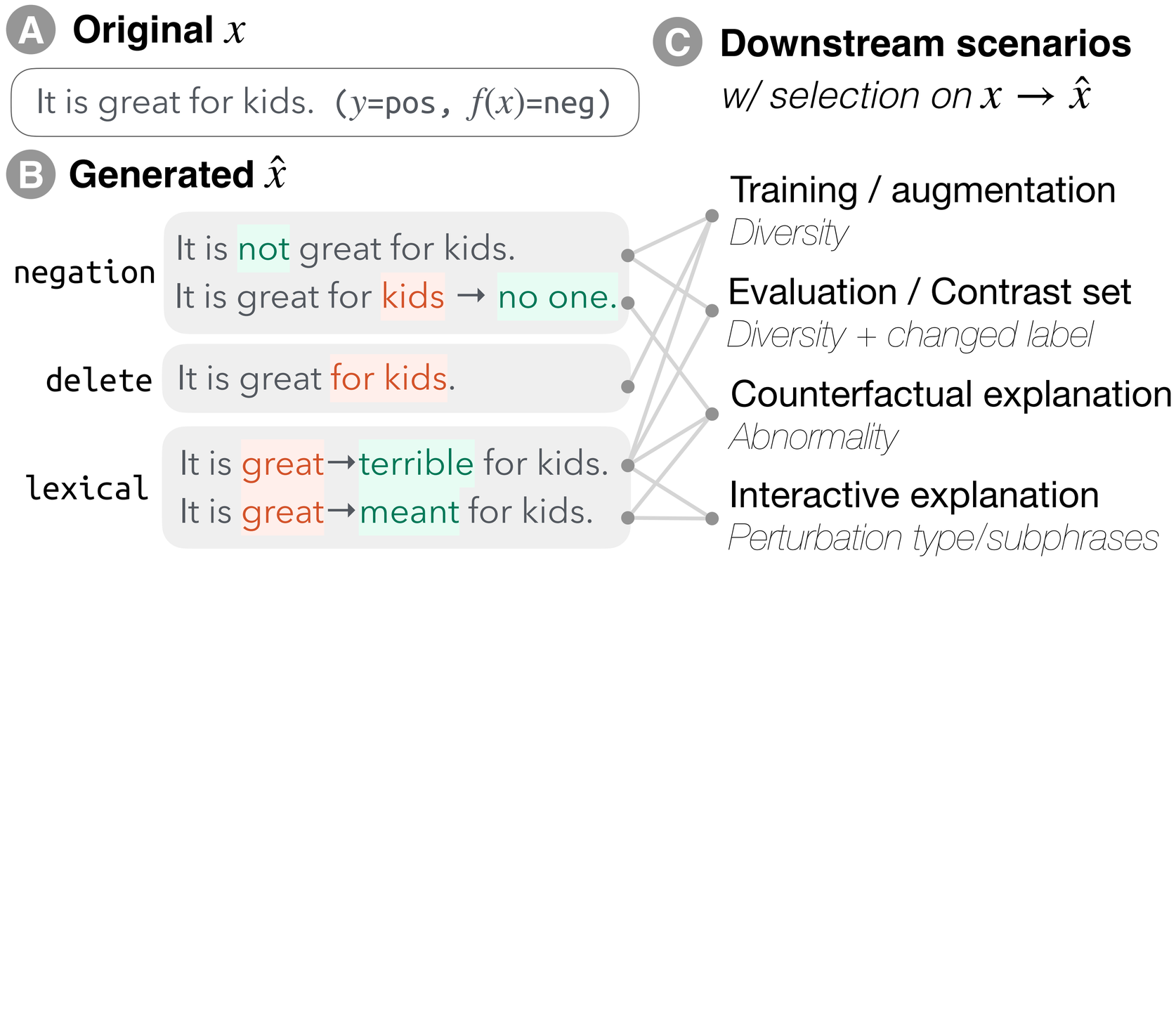
A Original x
It is great for kids. (y=pos, $f(x)$=neg)
B Generated $\hat{x}$
negation
It is not great for kids.
It is great for kids → no one.
delete
It is great for kids.
lexical
It is great→terrible for kids.
It is great→meant for kids.
C Downstream scenarios
w/ selection on $x \rightarrow \hat{x}$
Training / augmentation
Diversity
Evaluation / Contrast set
Diversity + changed label
Counterfactual explanation
Abnormality
Interactive explanation
Perturbation type/subphrases