COUNTERFACTUALS OF COUNTERFACTUALS

A BACK-TRANSLATION-INSPIRED APPROACH TO ANALYSE COUNTERFACTUAL EDITORS

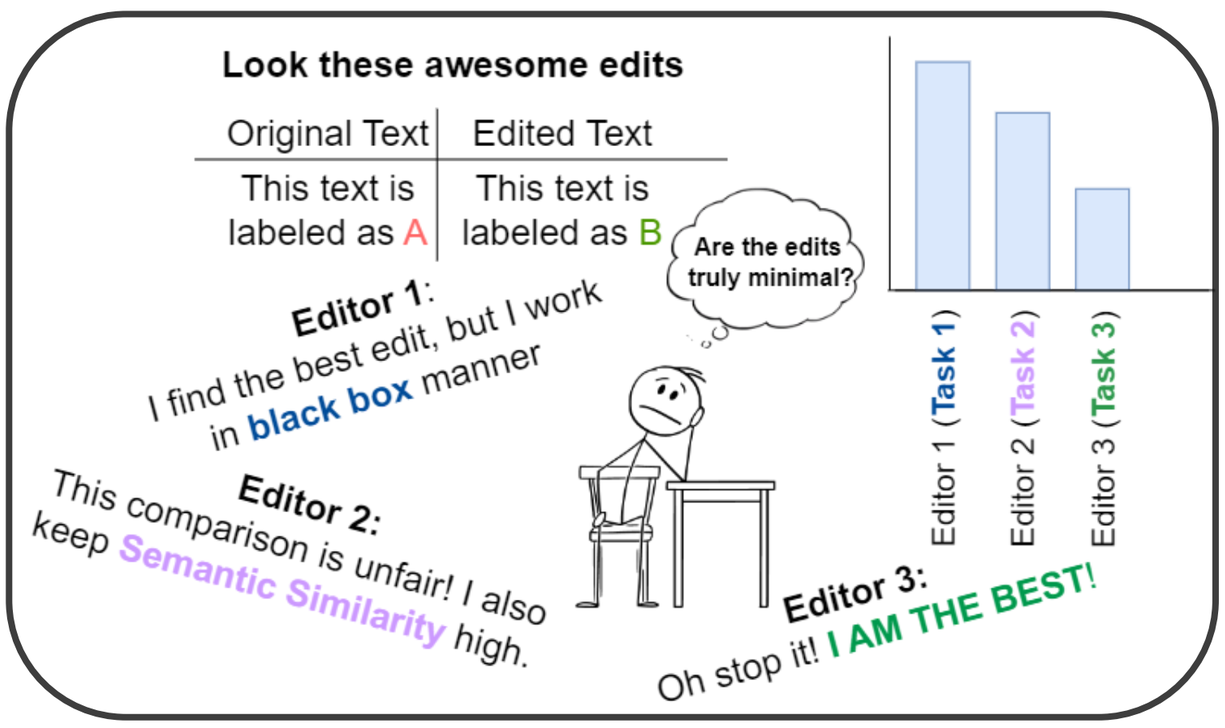
Giorgos Filandrianos[1], Edmund Dervakos[1], Orfeas Menis-Mastromichalakis[1], Chrysoula Zerva[2,3], Giorgos Stamou[1]
[1]National Technical University of Athens, [2]Instituto de Telecomunicações, [3]Instituto Superior Técnico & LUMLIS (Lisbon ELLIS Unit)
{geofila, eddiedervakos}@islab.ntua.gr, menorf@ails.ece.ntua.gr, chrysoula.zerva@tecnico.ulisboa.pt, gstam@cs.ntua.gr

TL;DR

We propose:

- **a method for evaluating counterfactual editors through iterative feedback steps.**
- **a novel metric inconsistency, setting an upper bound for the evaluation of the optimality of a counterfactual editor with respect to a specific criterion.**

EVALUATION
HOW IT IS TYPICALLY DONE



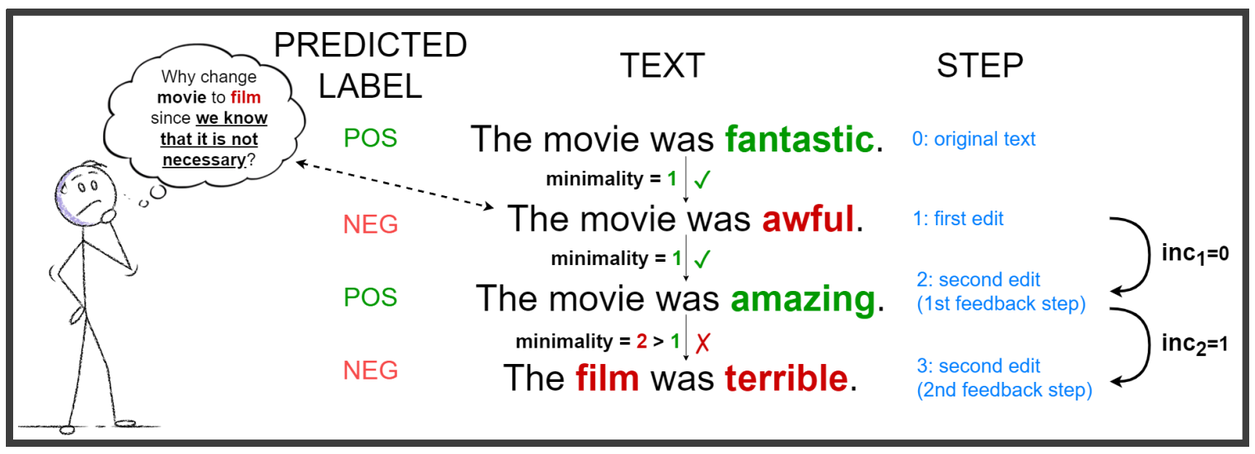
Problem: Lack of an ***ideal explanation*** to be used as ***ground truth*** means there is no way to know if a specific value of a metric is ***optimal*** (and generally good or bad).

INCONSISTENCY

By **recursively** using the system's output as input, we anticipate the result to be "***at least as good as***" the original input.
This is because the input is known, actionable, and feasible, serving as a "***lower bound***" for the generated edit and a **proxy for ground truth**.

$$\text{inc}(f, x) = \text{relu}[d(f(f(x)), f(x)) - d(f(x), x)]$$

$$\text{inc@}n(f, x) = \frac{1}{n}\sum_{i=0}^{n-1} \text{inc}(f_{i+1}(x), f_i(x))$$



EXPERIMENTS

We evaluate three editors by measuring inconsistency of minimality on two different datasets.

	IMDb			Newsgroups		
	MiCE	Polyjuice	TextFooler	MiCE	Polyjuice	TextFooler
inc@1 ↓	0.86	6.21	**0.01**	11.11	0.99	**0.04**
inc@2 ↓	5.95	4.65	**0.33**	7.97	1.29	**0.55**
inc@3 ↓	4.65	3.98	**0.36**	7.89	1.35	**0.46**
inc@5 ↓	4.87	2.9	**0.47**	6.92	1.3	**0.49**
inc@9 ↓	4.73	2.22	**0.49**	6.11	1.21	**0.46**

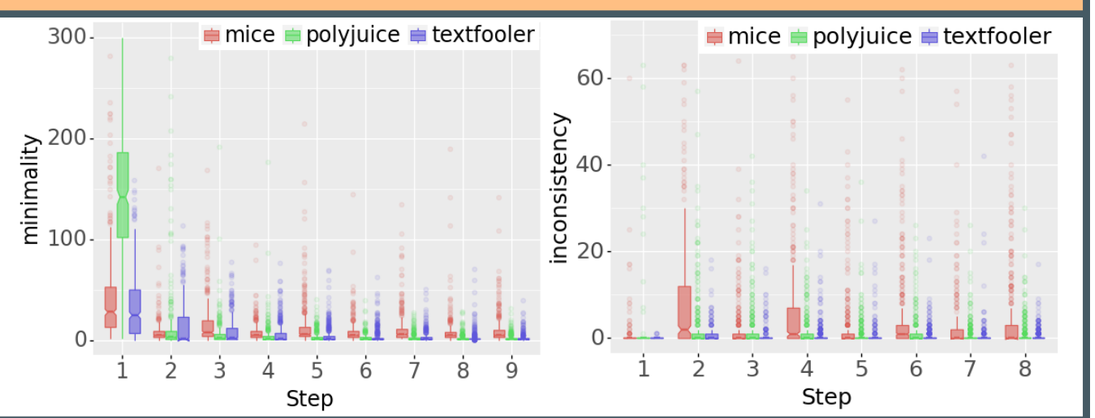
INTERPRETATION OF METRICS

▶ **Minimality** with value x → the editor is able to find counterexample with mean minimality = x. But do we know if the editor could be better?
▶ **Inconsistency of Minimality** with value x → the editor **misses a minimal counterfactual** solutions on average by at least x tokens.
▶ For example inc@2 = 5.95 → the editor **misses a minimal counterfactual** solutions on average by at least 5.95 tokens.

INSIGHTS

▶ MiCE and Polyjuice are less consistent than Textfooler probably due to the use of LLMs in the generation process, which are more sensitive to small perturbations that can alter their output.
▶ Polyjuice is more inconsistent for longer inputs and tends to make radical changes.
▶ MiCE has the lowest minimality but the highest inconsistency.

Minimality and inc@n, after each step of feedback and for each editor on the IMDb dataset.



MiCE tends to leave parts of the original text indicating the original sentiment (marked in bold) unchanged.

The biggest heroes, ***is one of the greatest movies ever****.* ***A good story, great actors and a brilliant ending*** *is what makes this film the ~~jumping start~~ absolute worst of the director Thomas Vinterberg's great ~~carrier~~ masterpiece.*

MAIN TAKEAWAYS

▶ **We can analyse different aspects of counterfactual editors and obtain an approximate ground truth by iteratively feeding back their output.**
▶ **The proposed inc@n metric can measure the consistency of editors, and can help gain new insights on their behaviour.**