ERASER: A Benchmark to Evaluate Rationalized NLP Models

The Evaluating Rationales And Simple English Reasoning

Alicja Gosiewska

Warszawa, 04.05.2020

Recent Publications

We actively publish academic research and present our findings at top nlp and computer vision conferences annually.

See all publications

The AI Economist: Improving Equality and Productivity with AI-Driven Tax Policies

Stephan Zheng, Alexander Trott, Sunil Srinivasa, Nikhil Naik, Melvin Gruesbeck, David C. Parkes, Richard Socher · arXiv

VD-BERT: A Unified Vision and Dialog Transformer with BERT

Yue Wang, Shafiq Joty, Michael R. Lyu, Irwin King, Caiming Xiong, Steven C. H. Hoi · arXiv

An investigation of phone-based subword units for end-to-end speech recognition

Weiran Wang, Yingbo Zhou, Caiming Xiong, Richard Socher · arXiv

ACL 2020, ICLR 2020, ITA 2020, AAAI 2020,
NeurIPS 2019, LVIS 2019, EMNLP 2019, ...

Benchmark

http://www.eraserbenchmark.com/

Eraser

The need for more interpretable models in NLP has become increasingly apparent in recent years. The Evaluating Rationales And Simple English Reasoning (ERASER) benchmark is intended to advance research in this area by providing a diverse set of NLP datasets that contain both document labels and snippets of text marked by annotators as supporting these.

Models that provide rationales supporting predictions can be evaluated using this benchmark using several metrics (*see below*) that aim to quantify different attributes of "interpretability". We do not privilege any one of these, or provide a single number to quantify performance, because we argue that the appropriate metric to gauge the quality of rationales will depend on the task and use case.

Tasks

Eraser
Website Link
Download

BoolQ
Website Link
Download

MultiRC
Website Link
Download

E-SNLI
Website Link
Download

CoS-E
Website Link
Download

Fever
Website Link
Download

Evidence Inference
Website Link
Download

Movies
Website Link
Download

Leaderboard

Click on a column to sort it. Up arrows denote sorting by ascending order, while down arrows denote descending order.

Dataset: *BoolQ* View Mode: *Dataset*

System	↑Prf.	↑IOU	↑Token F1	↑AUPRC	↑Comp.
Baseline/(BERT/GloVe)/Attention-weight rationales	0.471			0.525	
Baseline/(BERT/GloVe)/Gradient rationales	0.471			0.072	
Baseline/(BERT/GloVe)/LIME rationales	0.471			0.073	
Bert-to-Bert pipeline	0.544	0.052	0.134	0.340	
(Lehman et al., 2019) pipeline	0.411	0.050	0.127	0.248	

Rows per page: 5 1-5 of 5

Benchmark
http://www.eraserbenchmark.com/

Paper
https://arxiv.org/abs/1911.03429

ERASER: A Benchmark to Evaluate Rationalized NLP Models

Jay DeYoung[*Ψ], Sarthak Jain[*Ψ], Nazneen Fatema Rajani[*Φ], Eric Lehman[Ψ], Caiming Xiong[Φ], Richard Socher[Φ], and Byron C. Wallace[Ψ]

[*]Equal contribution.
[Ψ]Khoury College of Computer Sciences, Northeastern University
[Φ]Salesforce Research, Palo Alto, CA, 94301

Abstract

State-of-the-art models in NLP are now predominantly based on deep neural networks that are opaque in terms of how they come to make predictions. This limitation has increased interest in designing more interpretable deep models for NLP that reveal the 'reasoning' behind model outputs. But work in this direction has been conducted on different datasets and tasks with correspondingly unique aims and metrics; this makes it difficult to track progress. We propose the **E**valuating **R**ationales **A**nd **S**imple **E**nglish **R**easoning (**ERASER**) benchmark to advance research on interpretable models in NLP. This benchmark comprises multiple datasets and tasks for which human annotations of "rationales" (supporting evidence) have been collected. We propose several metrics that aim to capture how well the rationales provided by models align with human rationales, and also how *faithful* these rationales are (i.e., the degree to which provided rationales influenced the corresponding predictions). Our hope is that releasing this benchmark facilitates progress on designing more interpretable NLP systems. The benchmark, code, and documentation are available at https://www.eraserbenchmark.com/

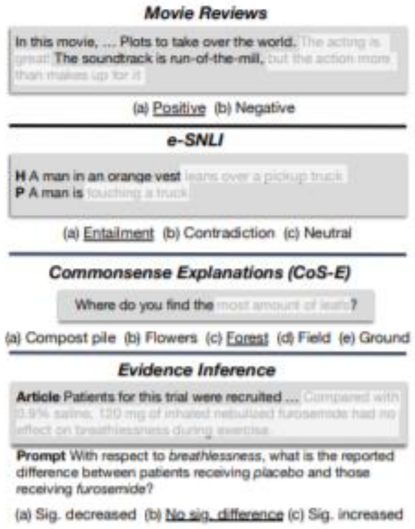


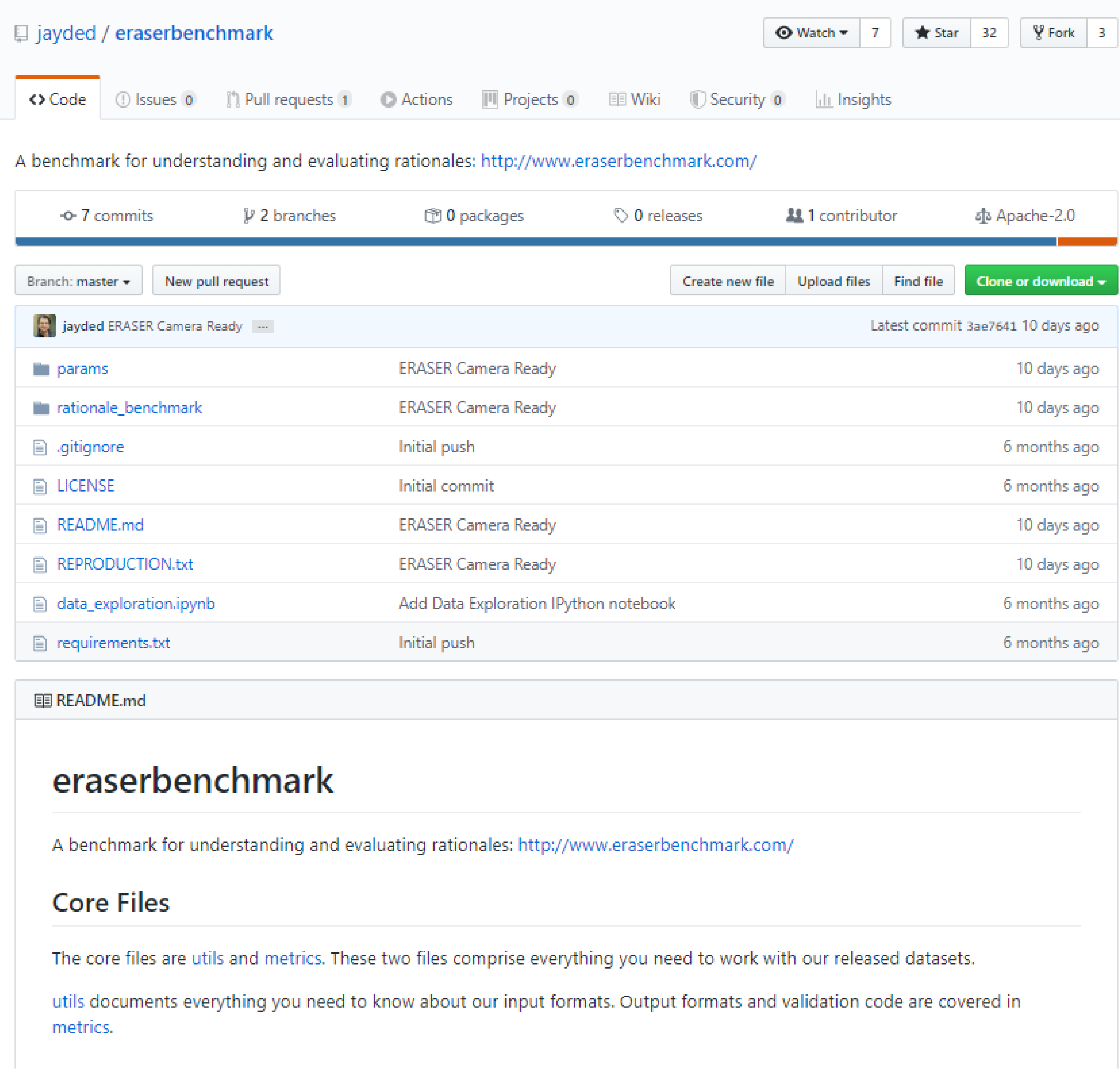
Figure 1: Examples of instances, labels, and rationales illustrative of four (out of seven) datasets included in ERASER. The 'erased' snippets are rationales.

In curating and releasing ERASER we take inspiration from the stickiness of the GLUE (Wang et al., 2019b) and SuperGLUE (Wang et al., 2019a) benchmarks for evaluating progress in natural language understanding tasks, which have driven rapid progress on models for general language repre-

Benchmark
http://www.eraserbenchmark.com/

Paper
https://arxiv.org/abs/1911.03429

GitHub
https://github.com/jayded/eraserbenchmark



Benchmark
http://www.eraserbenchmark.com/

Paper
https://arxiv.org/abs/1911.03429

GitHub
https://github.com/jayded/eraserbenchmark

Blog
https://blog.einstein.ai/eraser-a-benchmark-to-evaluate-rationalized-nlp-models/

ERASER: A Benchmark to Evaluate Rationalized NLP Models

By: Nazneen Rajani

Many NLP applications today deploy state-of-the-art deep neural networks that are essentially black-boxes. One of the goals of Explainable AI (XAI) is to have AI models reveal *why* and *how* they make their predictions so that these predictions are interpretable by a human. But work in this direction has been conducted on different datasets with correspondingly unique aims, and the inherent subjectivity in defining what constitutes 'interpretability' has resulted in no standard way to evaluate performance. Interpretability can mean multiple things depending on the task and context.

The **E**valuating **R**ationales **A**nd **S**imple **E**nglish **R**easoning (ERASER) benchmark is the first ever effort to unify and standardize NLP tasks with the goal of interpretability. Specifically, **we unify the definition of interpretability and metrics by using a standardized data collection and evaluation process for a suite of NLP tasks.**

This benchmark comprises 7 diverse NLP datasets and tasks for which we collected human annotations of explanations as supporting evidence for predictions. ERASER focuses on "rationales", that is, snippets of text extracted from the source document of the task that provides sufficient evidence for predicting the correct output. All the datasets included in ERASER are classification tasks including, sentiment analysis, Natural Language Inference, and Question Answering tasks, among others, with different number of labels and some have varying class labels. The figure below shows an example instance for 4 of the datasets and their corresponding classes as well as the rationales (erased) that support the predicted labels.

The **E**valuating **R**ationales **A**nd **S**imple **E**nglish **R**easoning (ERASER) benchmark is the first ever effort to unify and standardize NLP tasks with the goal of interpretability.

Consists of:
- 7 diverse NLP datasets and classification tasks including
- A suite of metrics to evaluate rationales

Movie Reviews

In this movie, … Plots to take over the world. The acting is great! The soundtrack is run-of-the-mill, but the action more than makes up for it

(a) Positive (b) Negative

e-SNLI

H A man in an orange vest leans over a pickup truck
P A man is touching a truck

(a) Entailment (b) Contradiction (c) Neutral

Commonsense Explanations (CoS-E)

Where do you find the most amount of leafs?

(a) Compost pile (b) Flowers (c) Forest (d) Field (e) Ground

Evidence Inference

Article Patients for this trial were recruited … Compared with 0.9% saline, 120 mg of inhaled nebulized furosemide had no effect on breathlessness during exercise.

Prompt With respect to *breathlessness*, what is the reported difference between patients receiving *placebo* and those receiving *furosemide*?

(a) Sig. decreased (b) No sig. difference (c) Sig. increased

Name	Size (train/dev/test)	Tokens	Comp?
Evidence Inference	7958 / 972 / 959	4761	◇
BoolQ	6363 / 1491 / 2817	3583	◇
Movie Reviews	1600 / 200 / 200	774	◆
FEVER	97957 / 6122 / 6111	327	✓
MultiRC	24029 / 3214 / 4848	303	✓
CoS-E	8733 / 1092 / 1092	28	✓
e-SNLI	911938 / 16449 / 16429	16	✓

Table 1: Overview of datasets in the ERASER benchmark. *Tokens* is the average number of tokens in each document. Comprehensive rationales mean that all supporting evidence is marked; ✓ denotes cases where this is (more or less) true by default; ◇, ◆ are datasets for which we have collected comprehensive rationales for either a subset or all of the test datasets, respectively. Additional information can be found in Appendix A.

Agreement with human rationales

Commonsense Explanations (CoS-E)

Where do you find the most amount of leafs?

(a) Compost pile (b) Flowers (c) Forest (d) Field (e) Ground

Discrete rationales:

First step

Intersection-Over-Union (IOU): for two spans, it is the size of the overlap of the tokens they cover divided by the size of their union.

A prediction is a match if it overlaps with any of the ground truth rationales by more than some threshold (here, 0.5).

Second step

These partial matches to calculate an **F1 score**.

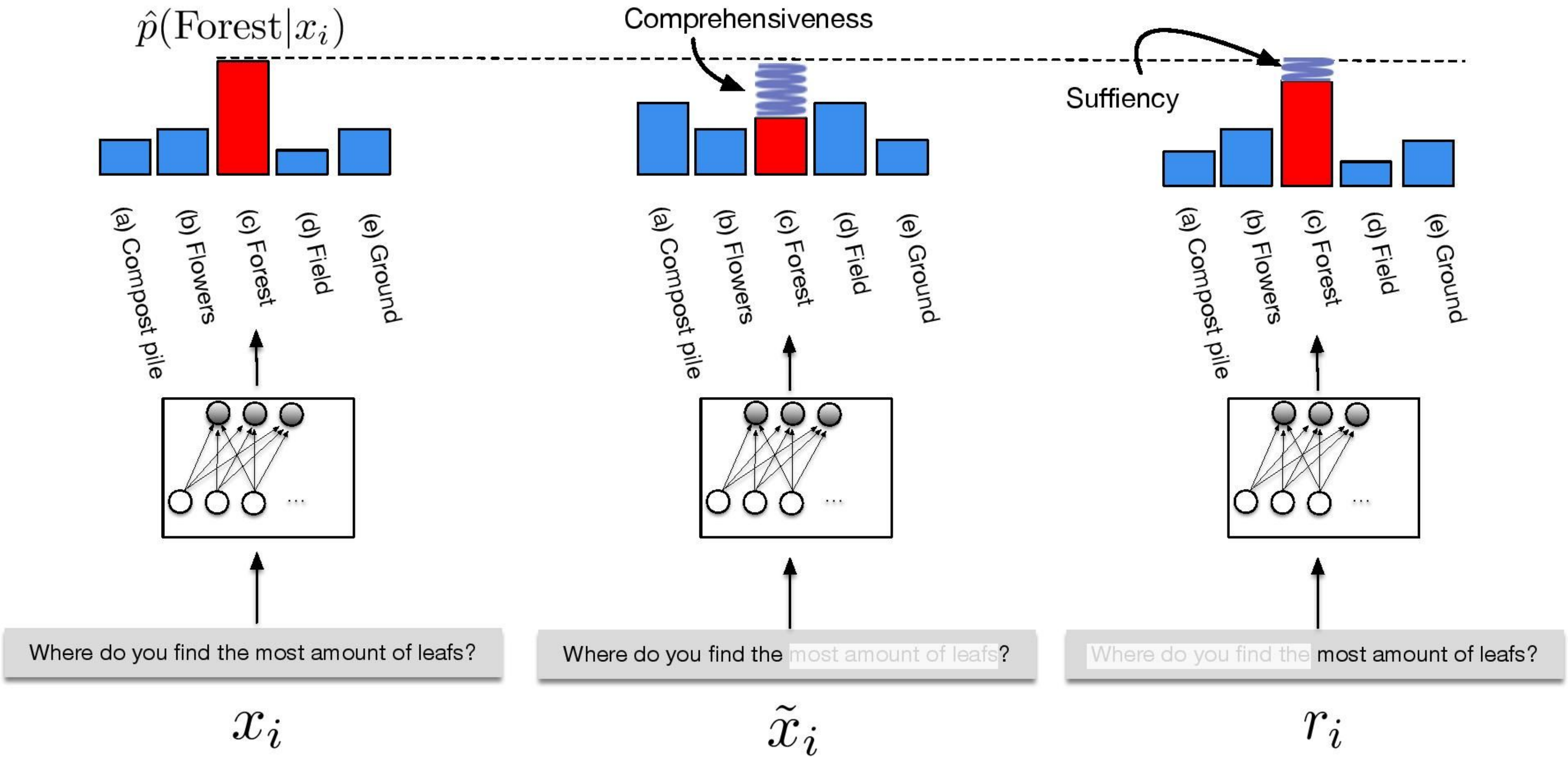
Continuous rationales:

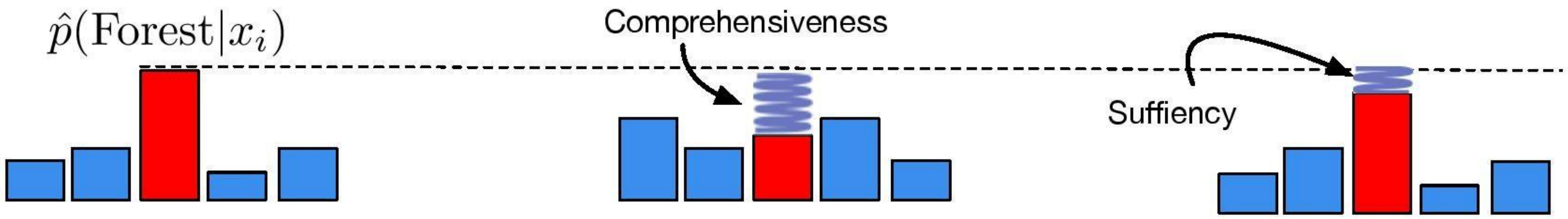
Area Under the Precision-Recall curve (AUPRC) is constructed by sweeping a threshold over token scores.

Sufficiency and comprehensiveness measures

Comprehensiveness is the measure of how much the model's prediction changes when it's not given the extracted rationale.

Sufficiency is the extent to which the extracted rationale was actually used by the model to make its prediction.





Discrete case

Comprehensiveness is the measure of how much the model's prediction changes when it's not given the extracted rationale.

x_i - an example
$\bar{x}_i$ - x_i with the predicted rationales r_i removed
$m(x_i)_j$ - an original prediction provided by a model m for the predicted class j

$$comprehensiveness = m(x_i)_j - m(\bar{x}_i)_j$$

Sufficiency is the extent to which the extracted rationale was actually used by the model to make its prediction.

$$sufficiency = m(x_i)_j - m(r_i)_j$$

Continuous case

comprehensiveness

Here we group tokens into k = 5 bins by grouping them into the top 1%, 5%, 10%, 20% and 50% of tokens, with respect to the corresponding importance score. We refer to these metrics as “Area Over the Perturbation Curve” (AOPC)

$$\frac{1}{|\mathcal{B}|+1}\left(\sum_{k=0}^{|\mathcal{B}|} m(x_i)_j - m(x_i \backslash r_{ik})_j\right)$$

	Perf.	AUPRC	Comp. ↑	Suff. ↓
Evidence Inference				
GloVe + LSTM - Attention	0.429	0.506	-0.002	-0.023
GloVe + LSTM - Gradient	0.429	0.016	0.046	-0.138
GloVe + LSTM - Lime	0.429	0.014	0.006	-0.128
GloVe + LSTM - Random	0.429	0.014	-0.001	-0.026
BoolQ				
GloVe + LSTM - Attention	0.471	0.525	0.010	0.022
GloVe + LSTM - Gradient	0.471	0.072	0.024	0.031
GloVe + LSTM - Lime	0.471	0.073	0.028	-0.154
GloVe + LSTM - Random	0.471	0.074	0.000	0.005
Movies				
BERT+LSTM - Attention	0.970	0.417	0.129	0.097
BERT+LSTM - Gradient	0.970	0.385	0.142	0.112
BERT+LSTM - Lime	0.970	0.280	0.187	0.093
BERT+LSTM - Random	0.970	0.259	0.058	0.330
FEVER				
BERT+LSTM - Attention	0.870	0.235	0.037	0.122
BERT+LSTM - Gradient	0.870	0.232	0.059	0.136
BERT+LSTM - Lime	0.870	0.291	0.212	0.014
BERT+LSTM - Random	0.870	0.244	0.034	0.122

	Perf.	AUPRC	Comp. ↑	Suff. ↓
MultiRC				
BERT+LSTM - Attention	0.655	0.244	0.036	0.052
BERT+LSTM - Gradient	0.655	0.224	0.077	0.064
BERT+LSTM - Lime	0.655	0.208	0.213	-0.079
BERT+LSTM - Random	0.655	0.186	0.029	0.081
CoS-E				
BERT+LSTM - Attention	0.487	0.606	0.080	0.217
BERT+LSTM - Gradient	0.487	0.585	0.124	0.226
BERT+LSTM - Lime	0.487	0.544	0.223	0.143
BERT+LSTM - Random	0.487	0.594	0.072	0.224
e-SNLI				
BERT+LSTM - Attention	0.960	0.395	0.105	0.583
BERT+LSTM - Gradient	0.960	0.416	0.180	0.472
BERT+LSTM - Lime	0.960	0.513	0.437	0.389
BERT+LSTM - Random	0.960	0.357	0.081	0.487

Table 4: Metrics for 'soft' scoring models. Perf. is accuracy (CoS-E) or F1 (others). Comprehensiveness and sufficiency are in terms of AOPC (Eq. 3). 'Random' assigns random scores to tokens to induce orderings; these are averages over 10 runs.

Benchmark
http://www.eraserbenchmark.com/

Paper
https://arxiv.org/abs/1911.03429

GitHub
https://github.com/jayded/eraserbenchmark

Blog
https://blog.einstein.ai/eraser-a-benchmark-to-evaluate-rationalized-nlp-models/