Adaptive Testing and Debugging of NLP Models

Emilia Wiśnios

MI2 Seminar, 29.05.2023

Problem

Finding and fixing bugs in LLMs remains a challenge

Example: Sentiment Analysis

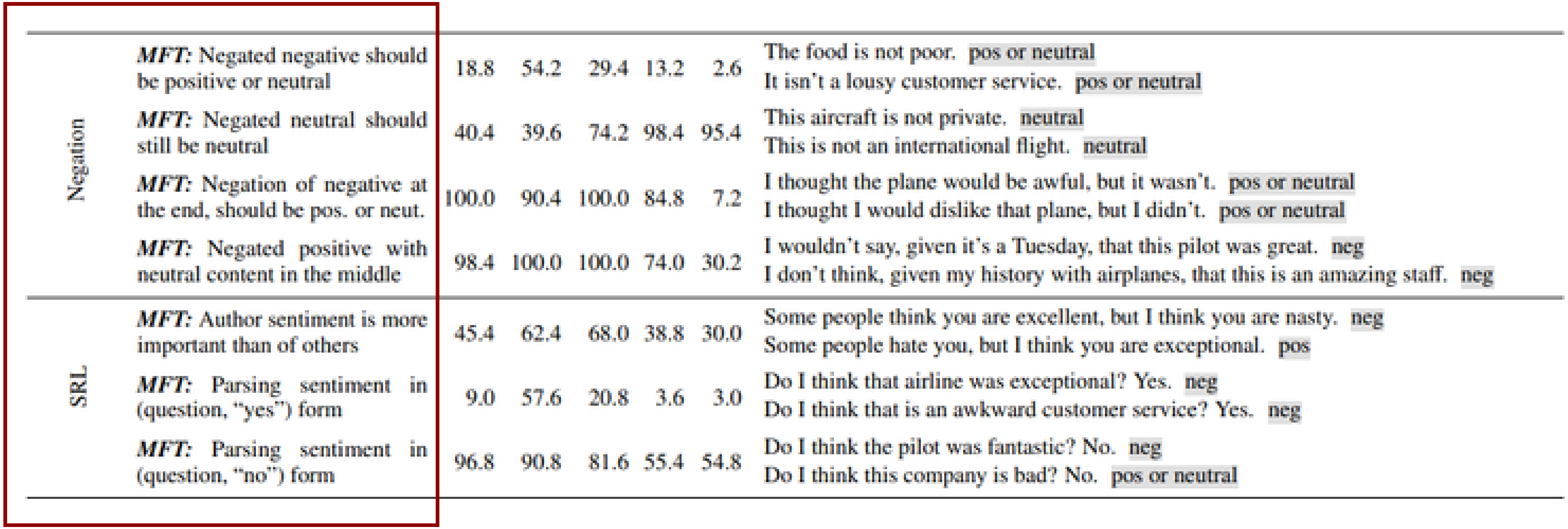
f(I am a black woman) ≠ negative ✓

f(I am a racial minority) ≠ negative ✗

Current approaches

CheckList

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList
Ribeiro et al.



Negation	*MFT:* Negated negative should be positive or neutral	18.8	54.2	29.4	13.2	2.6	The food is not poor. pos or neutral It isn't a lousy customer service. pos or neutral
	MFT: Negated neutral should still be neutral	40.4	39.6	74.2	98.4	95.4	This aircraft is not private. neutral This is not an international flight. neutral
	MFT: Negation of negative at the end, should be pos. or neut.	100.0	90.4	100.0	84.8	7.2	I thought the plane would be awful, but it wasn't. pos or neutral I thought I would dislike that plane, but I didn't. pos or neutral
	MFT: Negated positive with neutral content in the middle	98.4	100.0	100.0	74.0	30.2	I wouldn't say, given it's a Tuesday, that this pilot was great. neg I don't think, given my history with airplanes, that this is an amazing staff. neg
SRL	*MFT:* Author sentiment is more important than of others	45.4	62.4	68.0	38.8	30.0	Some people think you are excellent, but I think you are nasty. neg Some people hate you, but I think you are exceptional. pos
	MFT: Parsing sentiment in (question, "yes") form	9.0	57.6	20.8	3.6	3.0	Do I think that airline was exceptional? Yes. neg Do I think that is an awkward customer service? Yes. neg
	MFT: Parsing sentiment in (question, "no") form	96.8	90.8	81.6	55.4	54.8	Do I think the pilot was fantastic? No. neg Do I think this company is bad? No. pos or neutral

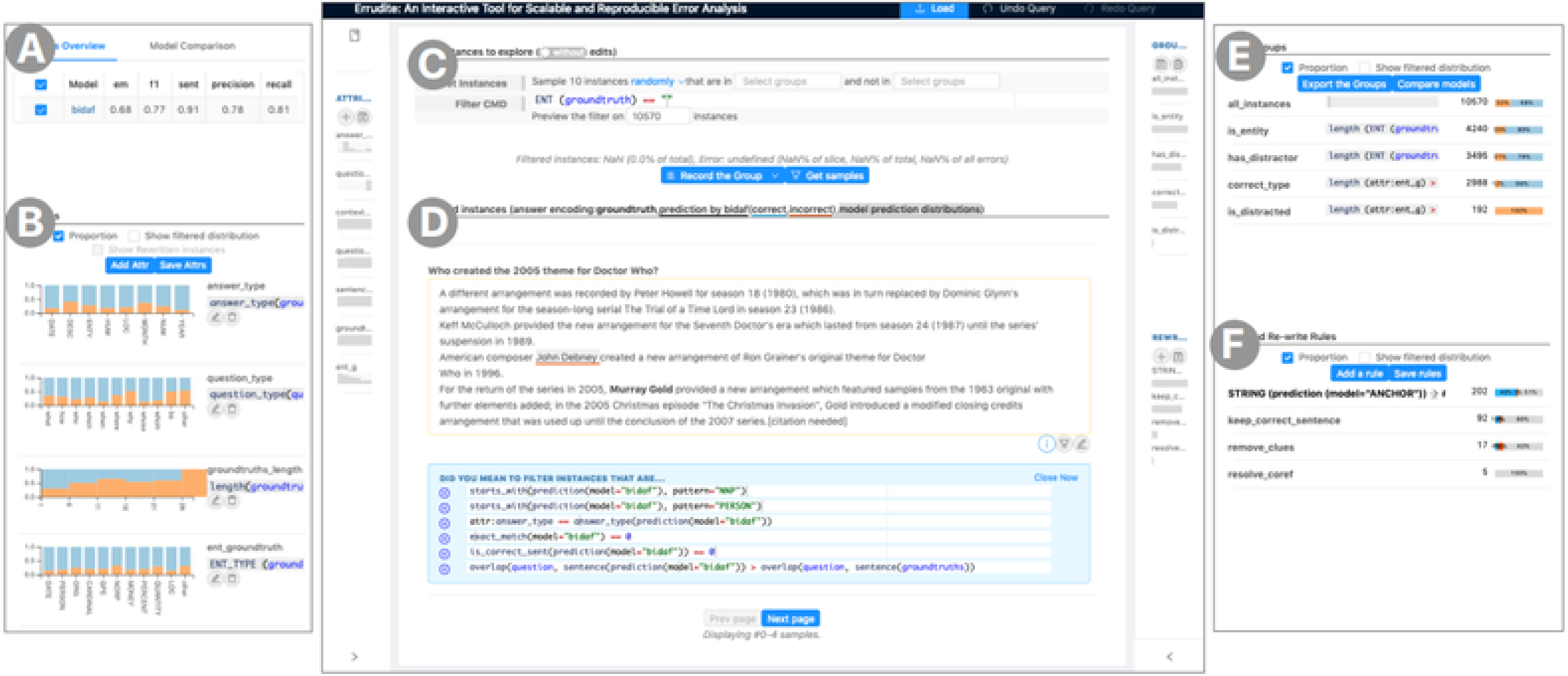
Current approaches

CheckList

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList
Ribeiro et al.

Errudite

Errudite: Scalable, Reproducible, and Testable Error Analysis
Wu et al.



Errudite: An Interactive Tool for Scalable and Reproducible Error Analysis

s Overview | Model Comparison

☑	Model	em	f1	sent	precision	recall
☑	bidaf	0.68	0.77	0.91	0.78	0.81

☑ Proportion □ Show filtered distribution
□ Show Rewritten Instances

answer_type
`answer_type(grou`

question_type
`question_type(qu`

groundtruths_length
`length(groundtru`

ent_groundtruth
`ENT_TYPE (ground`

ances to explore (edits)

t instances | Sample 10 instances randomly that are in Select groups and not in Select groups

Filter CMD | `ENT (groundtruth) == ""`

Preview the filter on 10570 instances

Filtered instances: NaN (0.0% of total), Error: undefined (NaN% of slice, NaN% of total, NaN% of all errors)

d instances (answer encoding groundtruth prediction by bidaf(correct incorrect) model prediction distributions)

Who created the 2005 theme for Doctor Who?

A different arrangement was recorded by Peter Howell for season 18 (1980), which was in turn replaced by Dominic Glynn's arrangement for the season-long serial The Trial of a Time Lord in season 23 (1986).
Keff McCulloch provided the new arrangement for the Seventh Doctor's era which lasted from season 24 (1987) until the series' suspension in 1989.
American composer John Debney created a new arrangement of Ron Grainer's original theme for Doctor Who in 1996.
For the return of the series in 2005, **Murray Gold** provided a new arrangement which featured samples from the 1963 original with further elements added; in the 2005 Christmas episode "The Christmas Invasion", Gold introduced a modified closing credits arrangement that was used up until the conclusion of the 2007 series.[citation needed]

DID YOU MEAN TO FILTER INSTANCES THAT ARE...

```
starts_with(prediction(model="bidaf"), pattern="NNP")
starts_with(prediction(model="bidaf"), pattern="PERSON")
attr:answer_type == answer_type(prediction(model="bidaf"))
exact_match(model="bidaf") == 0
is_correct_sent(prediction(model="bidaf")) == 0
overlap(question, sentence(prediction(model="bidaf"))) > overlap(question, sentence(groundtruths))
```

Displaying #0–4 samples.

oups

☑ Proportion □ Show filtered distribution

all_instances		10570	32% 68%
is_entity	`length (ENT (groundtr`	4240	15% 85%
has_distractor	`length (ENT (groundtr`	3495	21% 79%
correct_type	`length (attr:ent_g) >`	2988	6% 94%
is_distracted	`length (attr:ent_g) >`	192	100%

d Re-write Rules

☑ Proportion □ Show filtered distribution

STRING (prediction (model="ANCHOR")) -> #	202	46% 6% 51%
keep_correct_sentence	92	85%
remove_clues	17	80%
resolve_coref	5	100%

Current approaches

CheckList

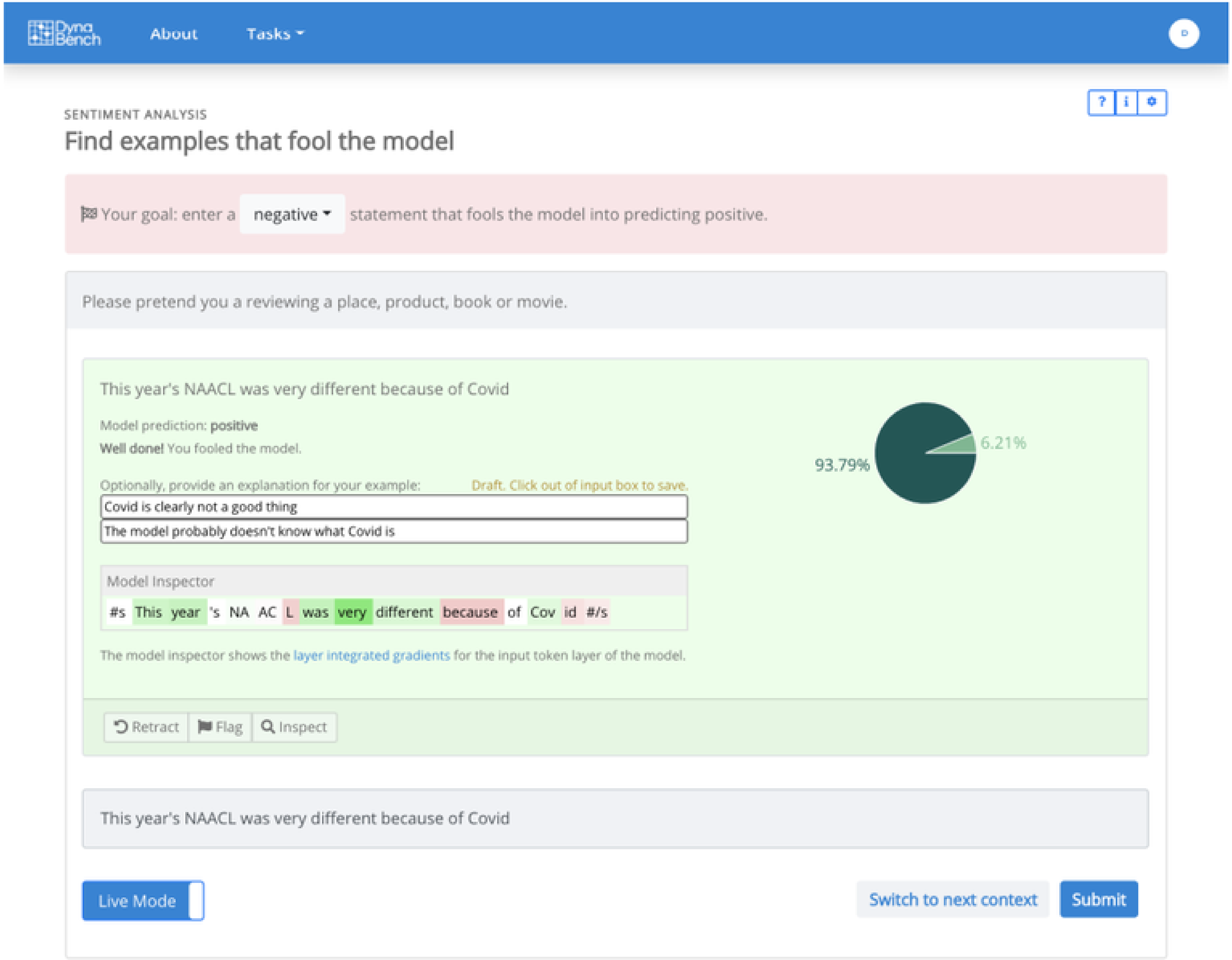
Beyond Accuracy: Behavioral Testing of NLP Models with CheckList
Ribeiro et al.

Errudite

Errudite: Scalable, Reproducible, and Testable Error Analysis
Wu et al.

Dynabench

Dynabench: Rethinking Benchmarking in NLP
Kiela et al.



SENTIMENT ANALYSIS

Find examples that fool the model

Your goal: enter a negative statement that fools the model into predicting positive.

Please pretend you a reviewing a place, product, book or movie.

This year's NAACL was very different because of Covid

Model prediction: **positive**

Well done! You fooled the model.

Optionally, provide an explanation for your example: Draft. Click out of input box to save.

Covid is clearly not a good thing

The model probably doesn't know what Covid is

Model Inspector

#s This year 's NA AC L was very different because of Cov id #/s

The model inspector shows the layer integrated gradients for the input token layer of the model.

This year's NAACL was very different because of Covid

Automatic approaches

Perturbations

Automatic Adversarial Examples

Unguided Data Augmentation

Perturbations

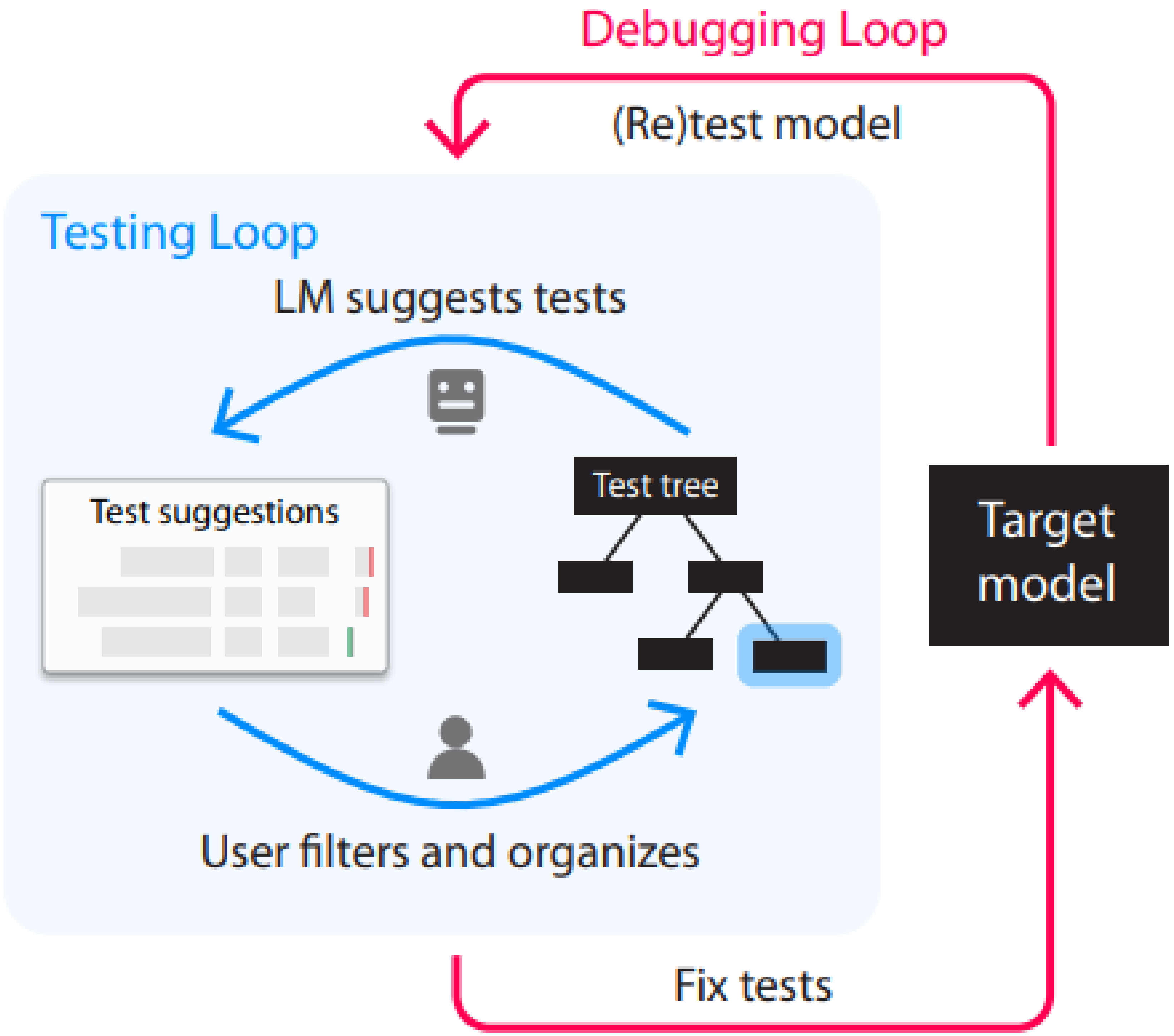
Automatic Adversarial Examples

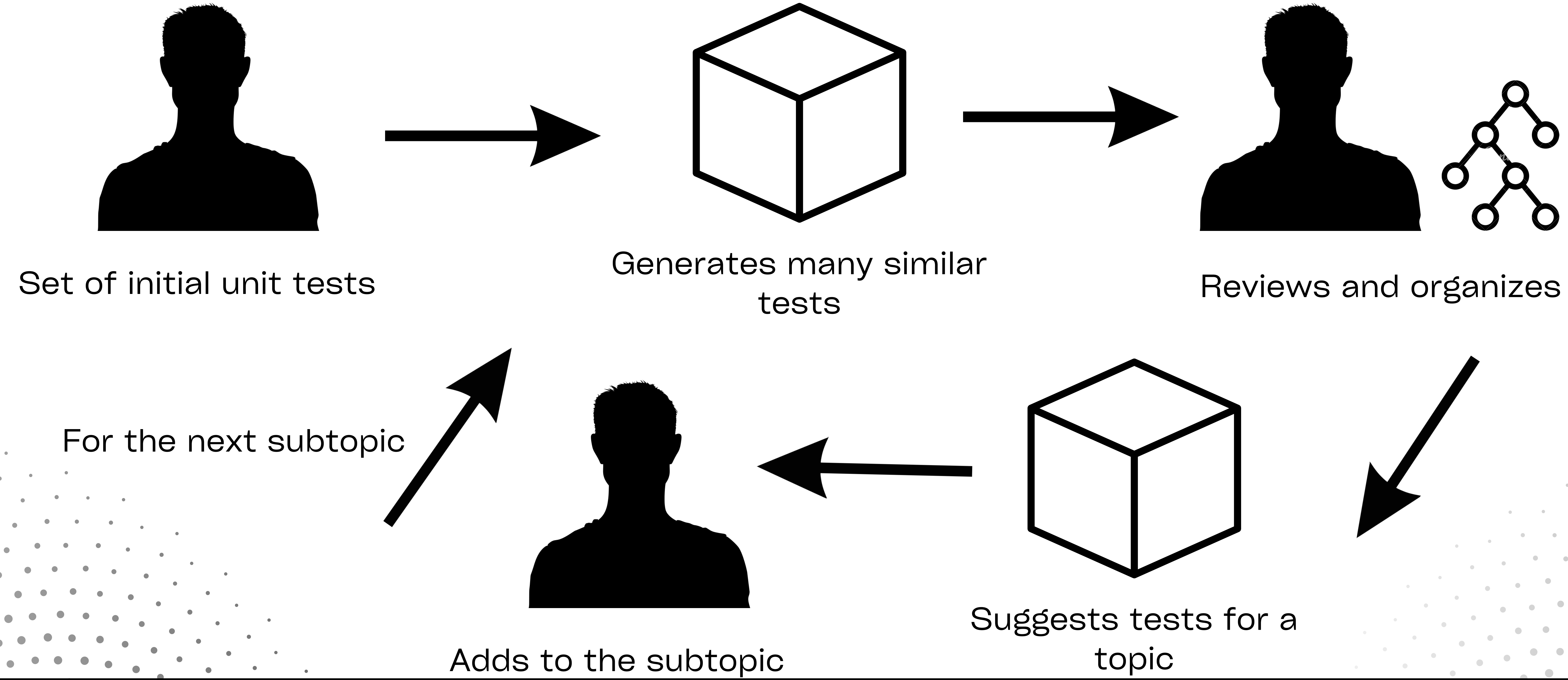
Unguided Data Augmentation

RESTRICTED TO SPECIFIC KIND OF PROBLEMS

Automatic approaches

AdaTest


Debugging Loop
(Re)test model
Testing Loop
LM suggests tests
Test suggestions
Test tree
Target model
User filters and organizes
Fix tests



Set of initial unit tests

Generates many similar tests

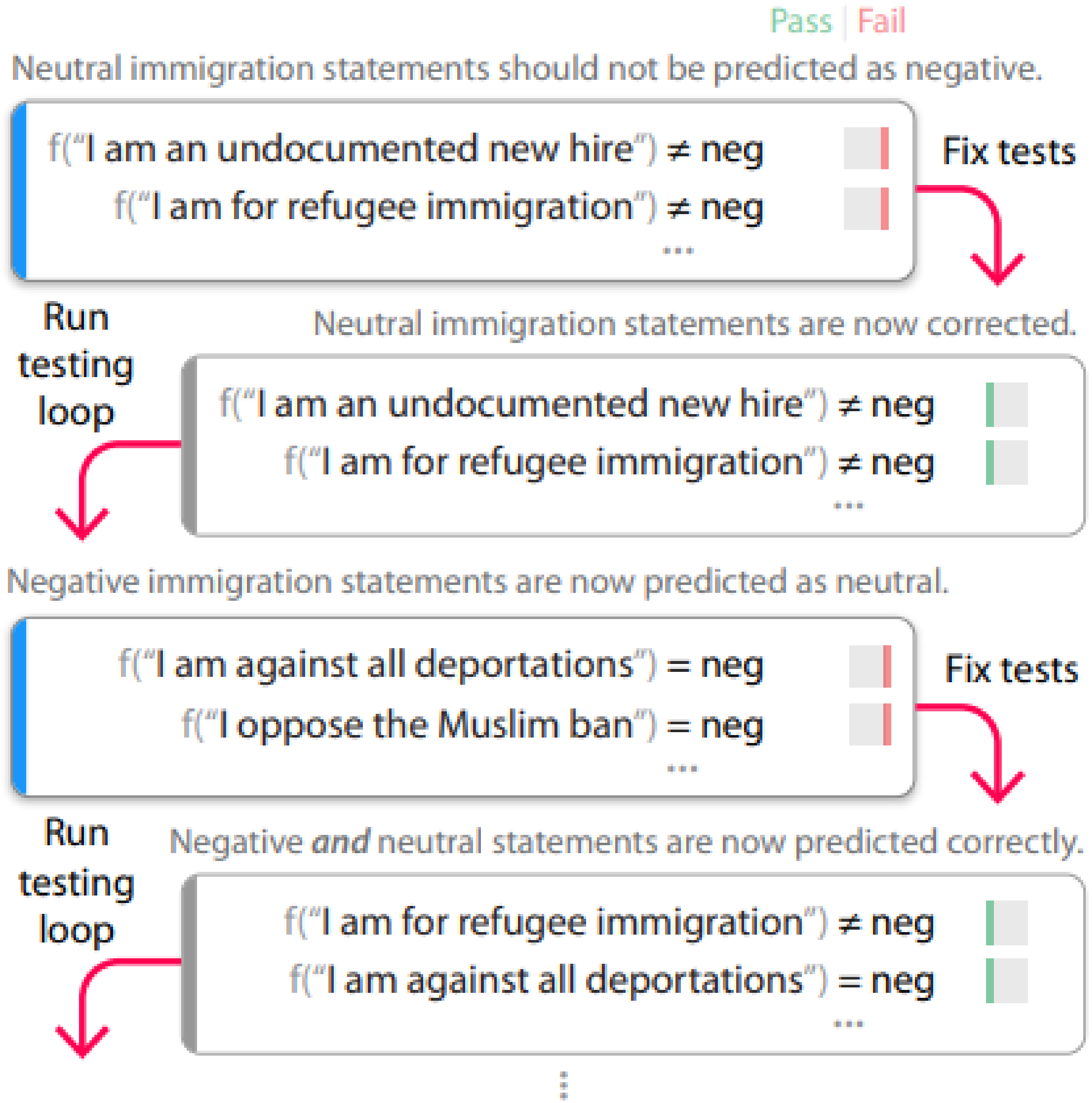
Reviews and organizes

Suggests tests for a topic

Adds to the subtopic

For the next subtopic

Testing Loop



Pass | Fail

Neutral immigration statements should not be predicted as negative.

f("I am an undocumented new hire") ≠ neg 🟥

f("I am for refugee immigration") ≠ neg 🟥

...

Fix tests

Run testing loop

Neutral immigration statements are now corrected.

f("I am an undocumented new hire") ≠ neg 🟩

f("I am for refugee immigration") ≠ neg 🟩

...

Negative immigration statements are now predicted as neutral.

f("I am against all deportations") = neg 🟥

f("I oppose the Muslim ban") = neg 🟥

...

Fix tests

Run testing loop

Negative *and* neutral statements are now predicted correctly.

f("I am for refugee immigration") ≠ neg 🟩

f("I am against all deportations") = neg 🟩

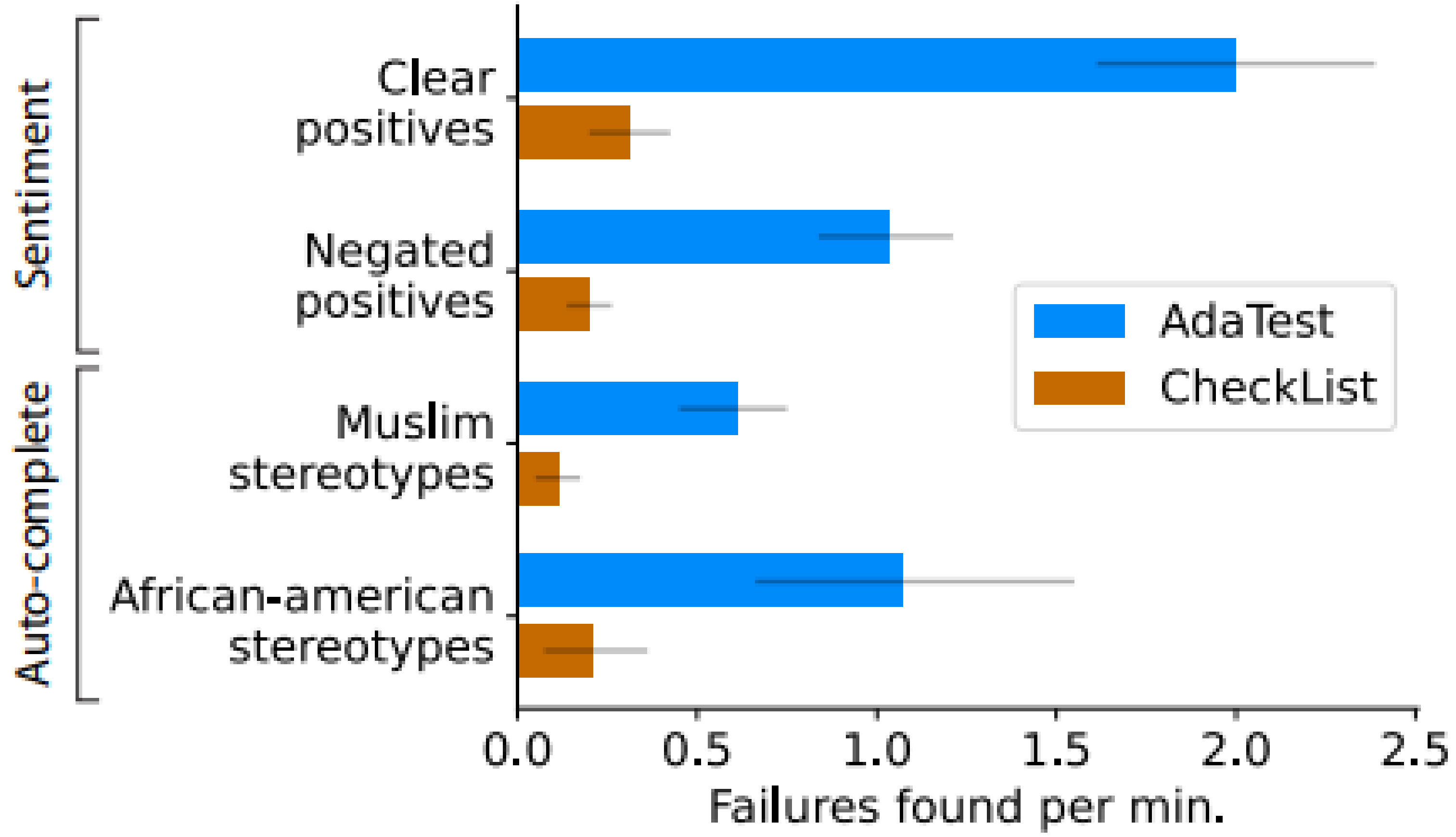
...

⋮

Debugging loop

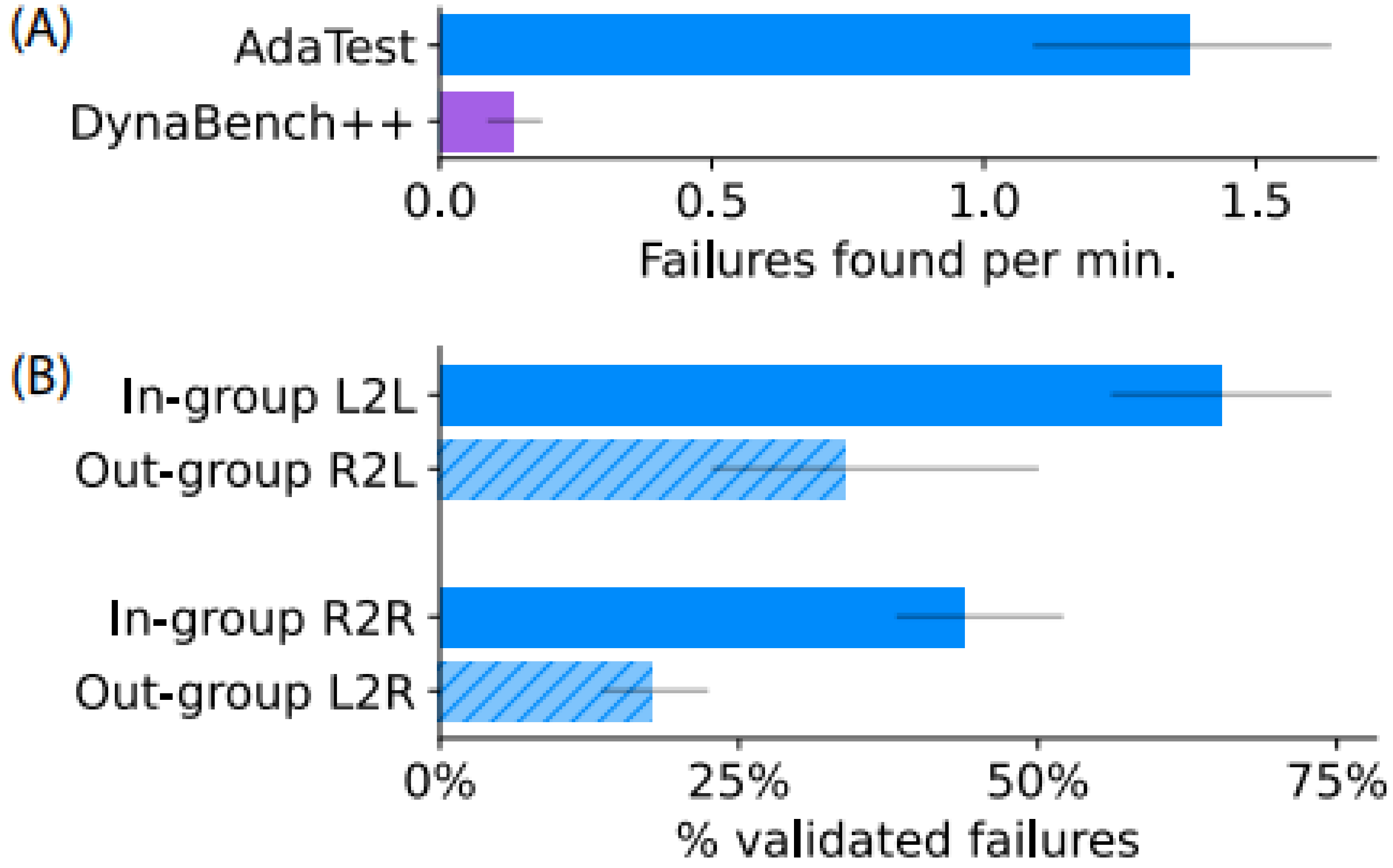
Evaluation

Expert evaluation



Evaluation

Non-expert evaluation



Case studies

Non-expert testing of non-classification models

Text to video matching

Task detection

Adaptive Testing and Debugging of NLP Models