Towards Debiasing Fact Verification Models

Tal Schuster*, Darsh J Shah*, Serene Yeo,

Daniel Filizzola, Enrico Santus, Regina Barzilay

Task: Fact verification

Is the following claim true?

Claim:

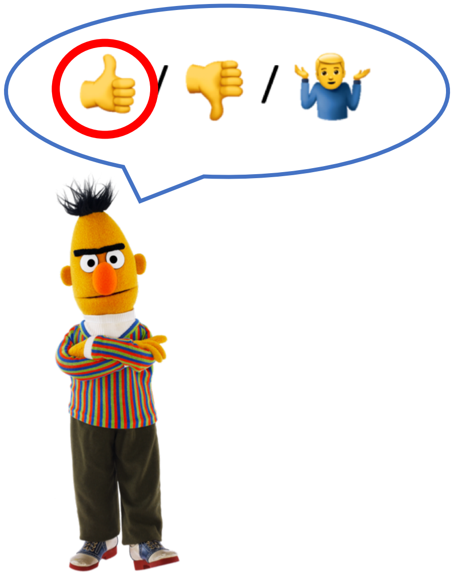
Trevor Griffiths was born on April 4, 1935

Task: Fact verification

Bert can answer correctly (using only the claim)

Claim:

Trevor Griffiths was born on April 4, 1935



Bert

Task: Fact verification

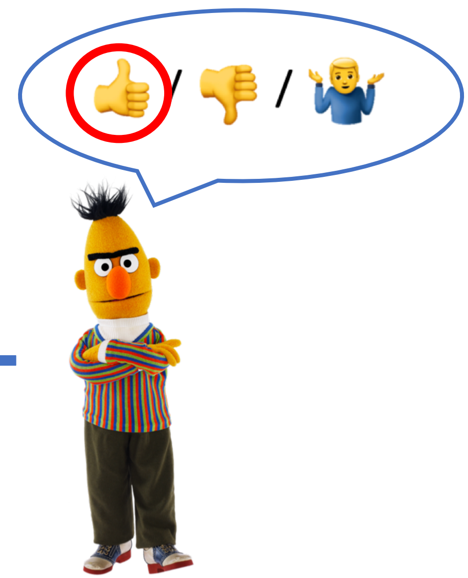
Shouldn’t it rely on evidence?

Claim:

Trevor Griffiths was born on April 4, 1935

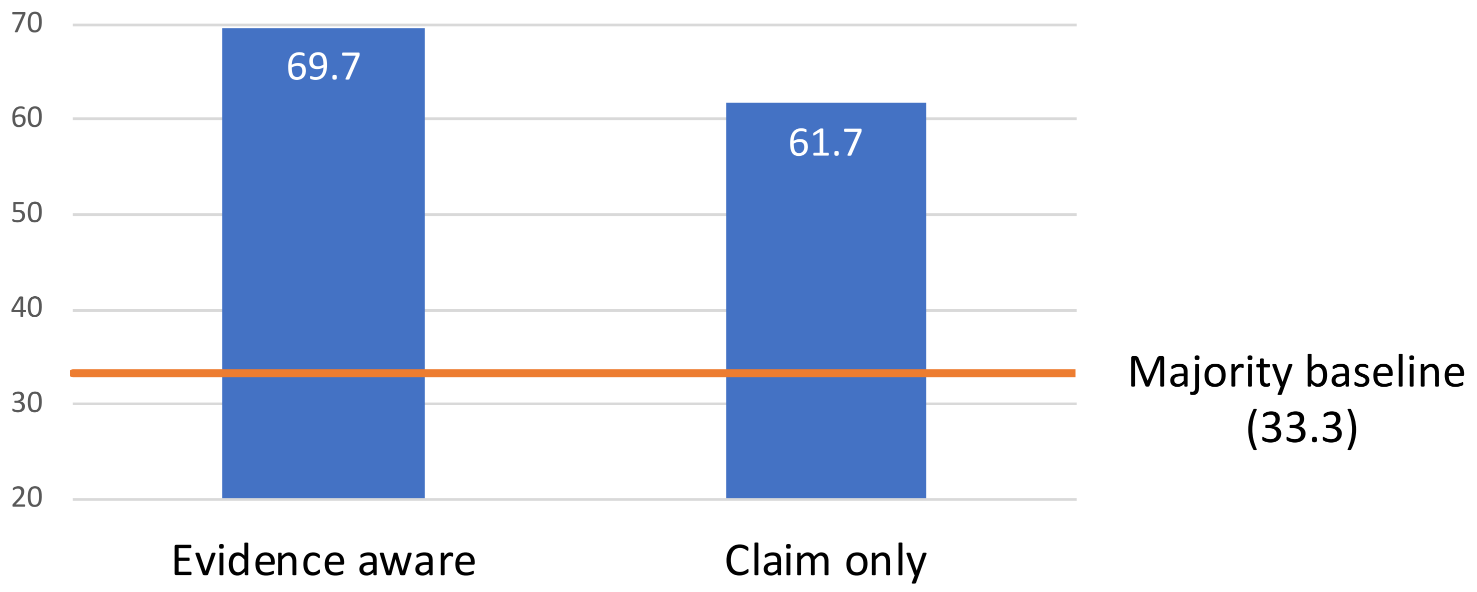
Evidence:

Trevor Griffiths (born 4 April 1935), is an English dramatist.



Bert

Performance on the FEVER dataset



Why does the claim-only model perform so well?

Possibility 1

World knowledge captured in the pretraining process

1. pretraining

WIKIPEDIA

2. fact-checking

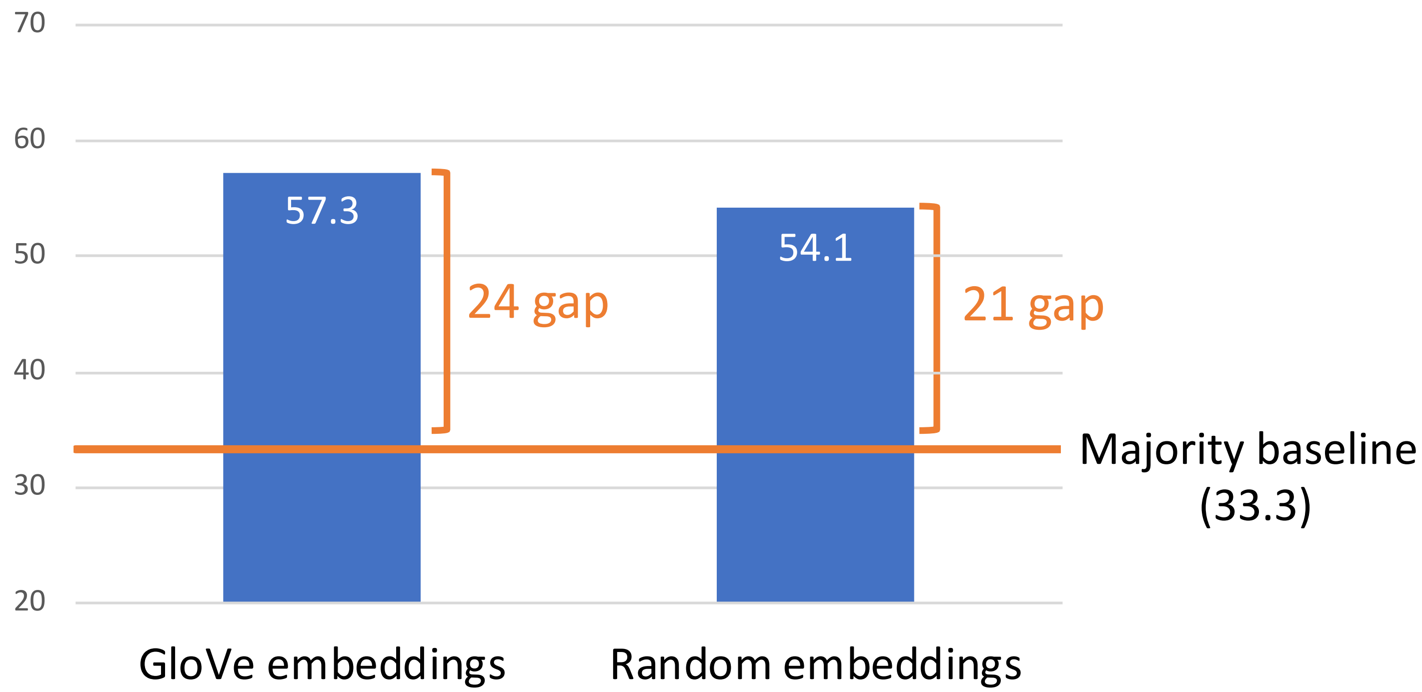
Claim based on Wikipedia

👍 / 👎 / 🤷‍♂️

can leak the truthfulness of claims

Without pretrained embeddings

Random Embeddings still perform far above the baseline



claim-only InferSent model (Conneau et al., Poliak et al.)

Possibility 2

Give-away phrases in the claims

Give-away phrases in the claims

<u>A claim-only model should fail:</u>

- Magic Johnson did not play for the Lakers.
- There has been at least one windstorm in Stanley Park.
- All About Eve won an award for Best Picture.
- The New England Patriots failed to reach seven Super Bowls.
- Quinoa did not originate in South America.

Give-away phrases in the claims

But the model can ”cheat”:

- 👎 Magic Johnson did not play for the Lakers.
- 👍 There has been at least one windstorm in Stanley Park.
- 👍 All About Eve won an award for Best Picture.
- 👎 The New England Patriots failed to reach seven Super Bowls.
- 👎 Quinoa did not originate in South America.

Give-away phrases in the claims

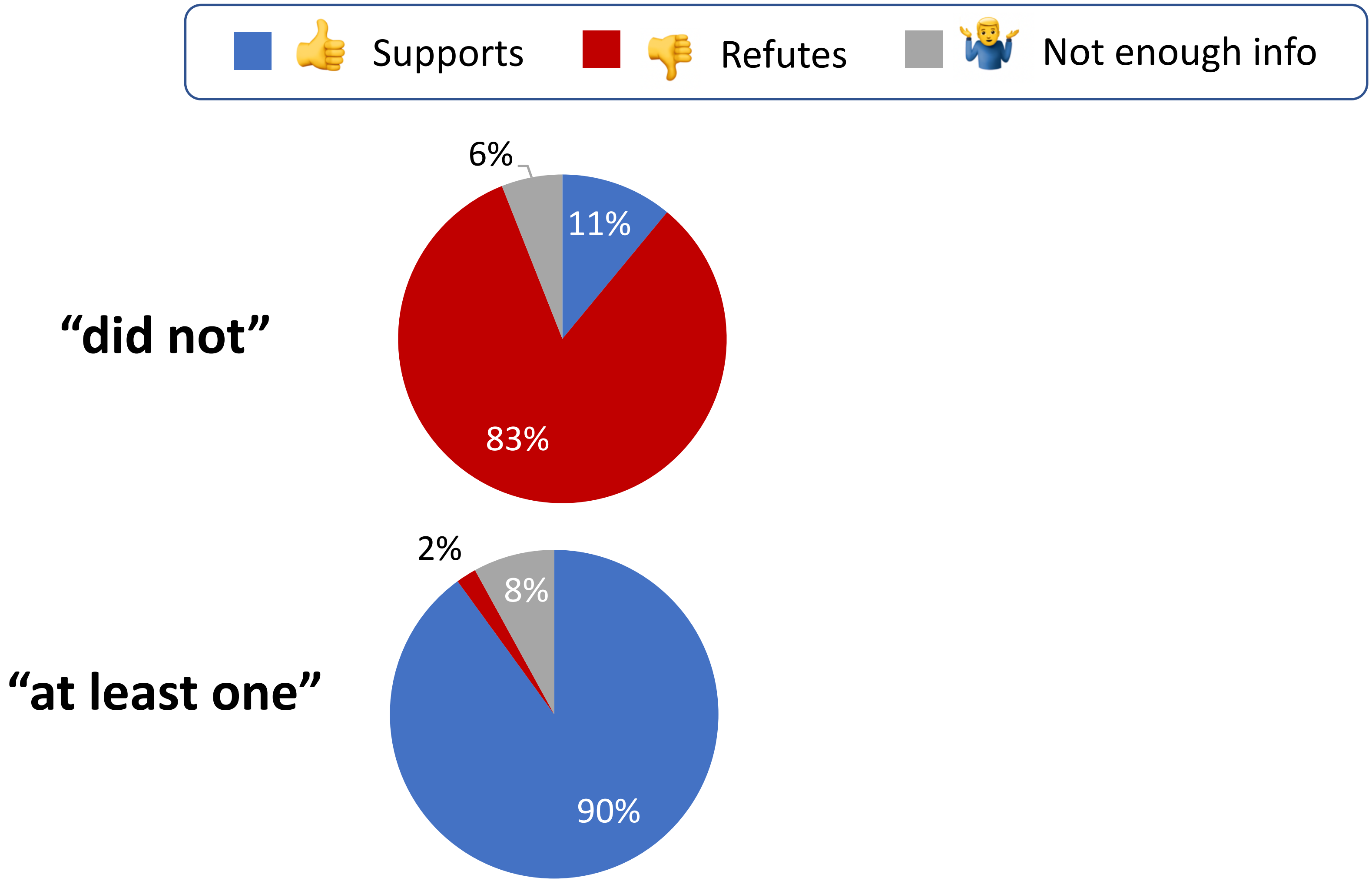
- Probability of a claim having label l if it contains phrase w_j:

$$p(l|w_j) = \frac{count(l, w_j)}{count(w_j)} = \frac{\sum_{i=1}^{n} \mathbb{1}_{\left[w_j^{(i)}\right]} \cdot \mathbb{1}_{\left[y^{(i)}=l\right]}}{\sum_{i=1}^{n} \mathbb{1}_{\left[w_j^{(i)}\right]}}$$

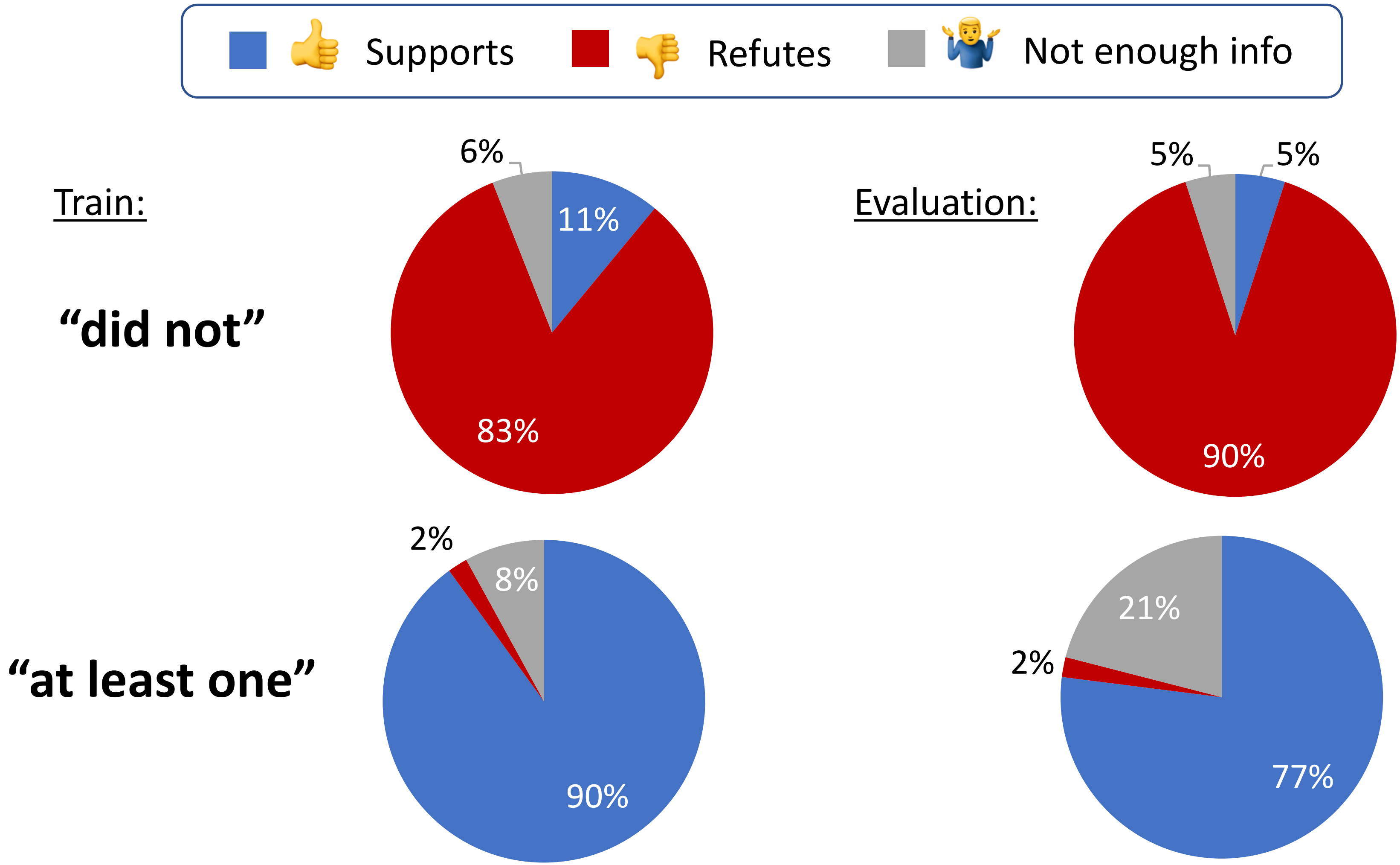
- Local Mutual Information - phrases that create the strongest bias:

$$LMI(w_j, l) = p(w_j, l) \cdot \underbrace{\log\left(\frac{p(l|w_j)}{p(l)}\right)}_{\text{PMI}}$$

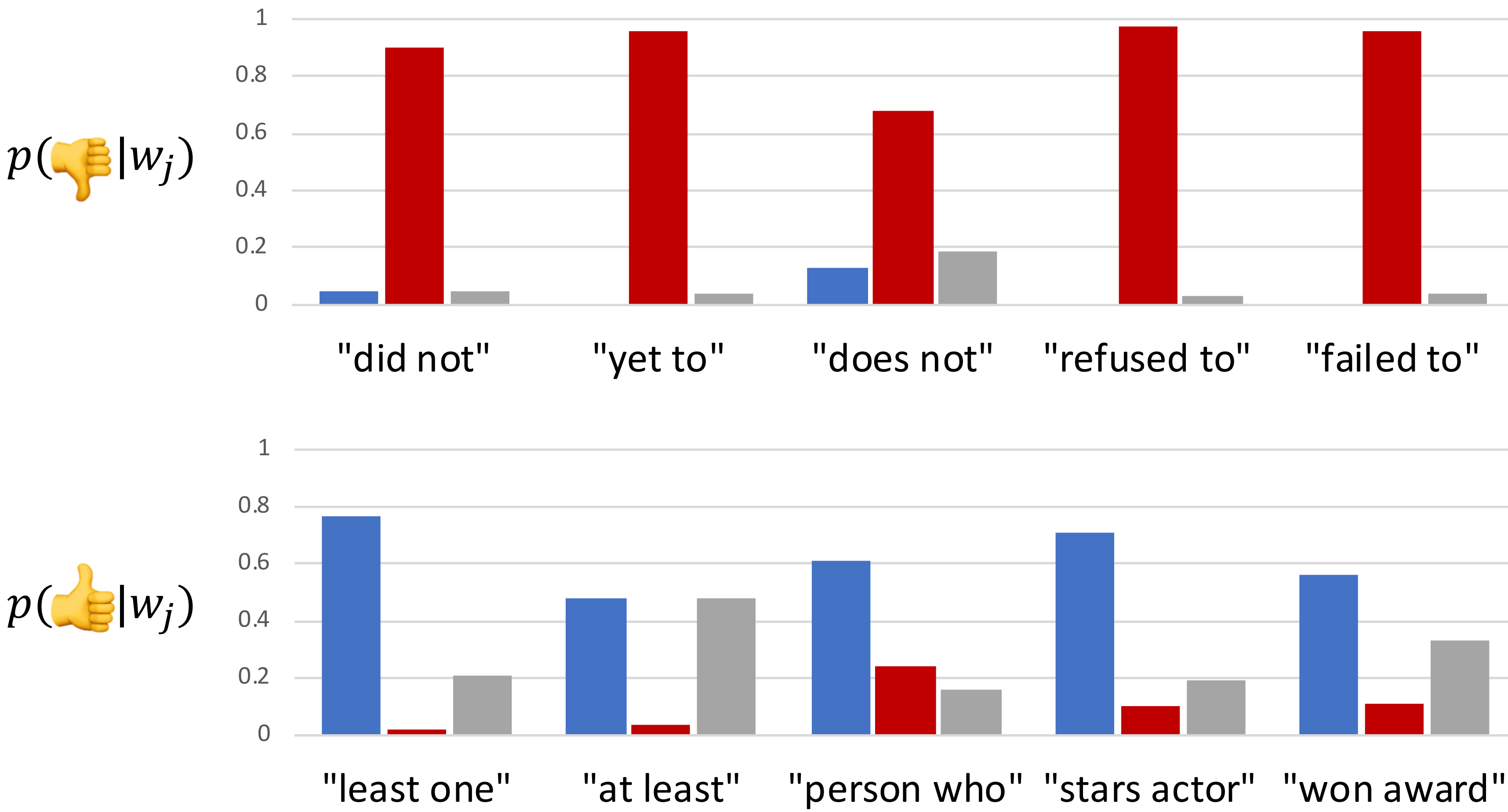
Example $p(l|w_j)$ for top LMI phrases



Bias reappears in the evaluation set



$p(l|w_j)$ on evaluation set for top LMI phrases by training set



Give-away phrases are the main culprits

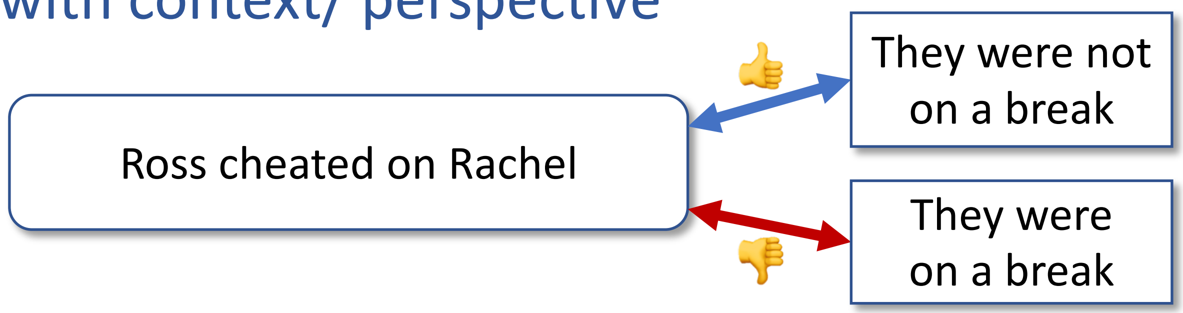
1. World knowledge captured in pretraining 
2. Give-away phrases in claims 

Goal: Evidence based predictions

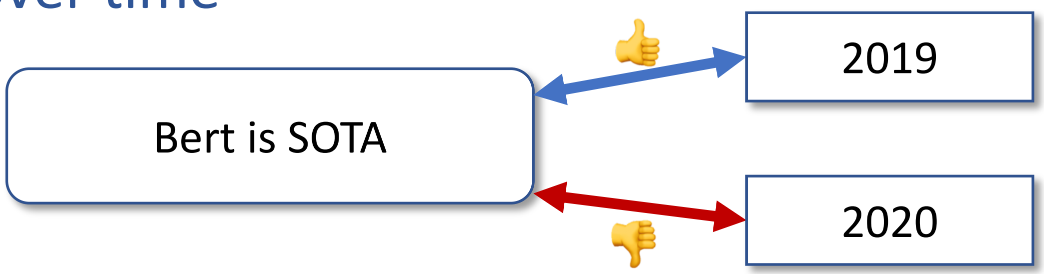
Claim-only should be inadequate

A claim's truthfulness might:

Vary with context/ perspective



Change over time



Debiasing the evaluation

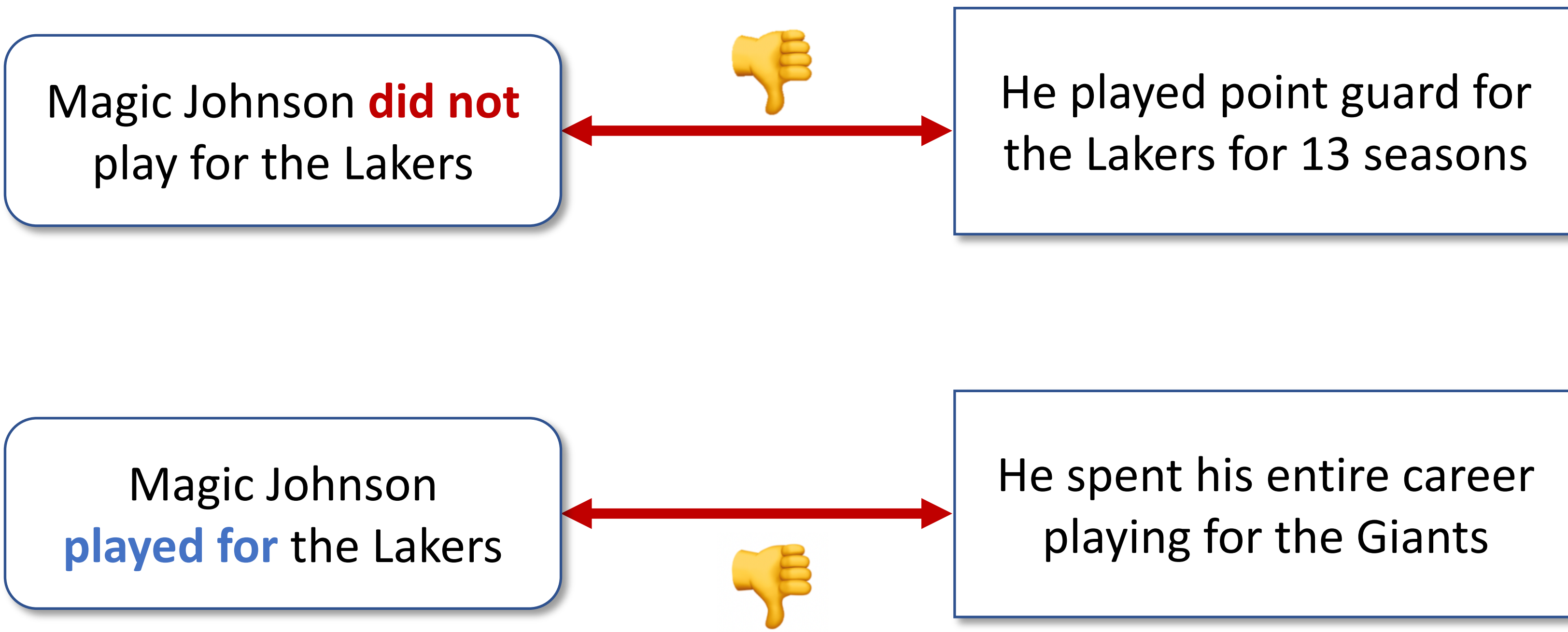
By creating a symmetric dataset

Creating a symmetric unbiased dataset

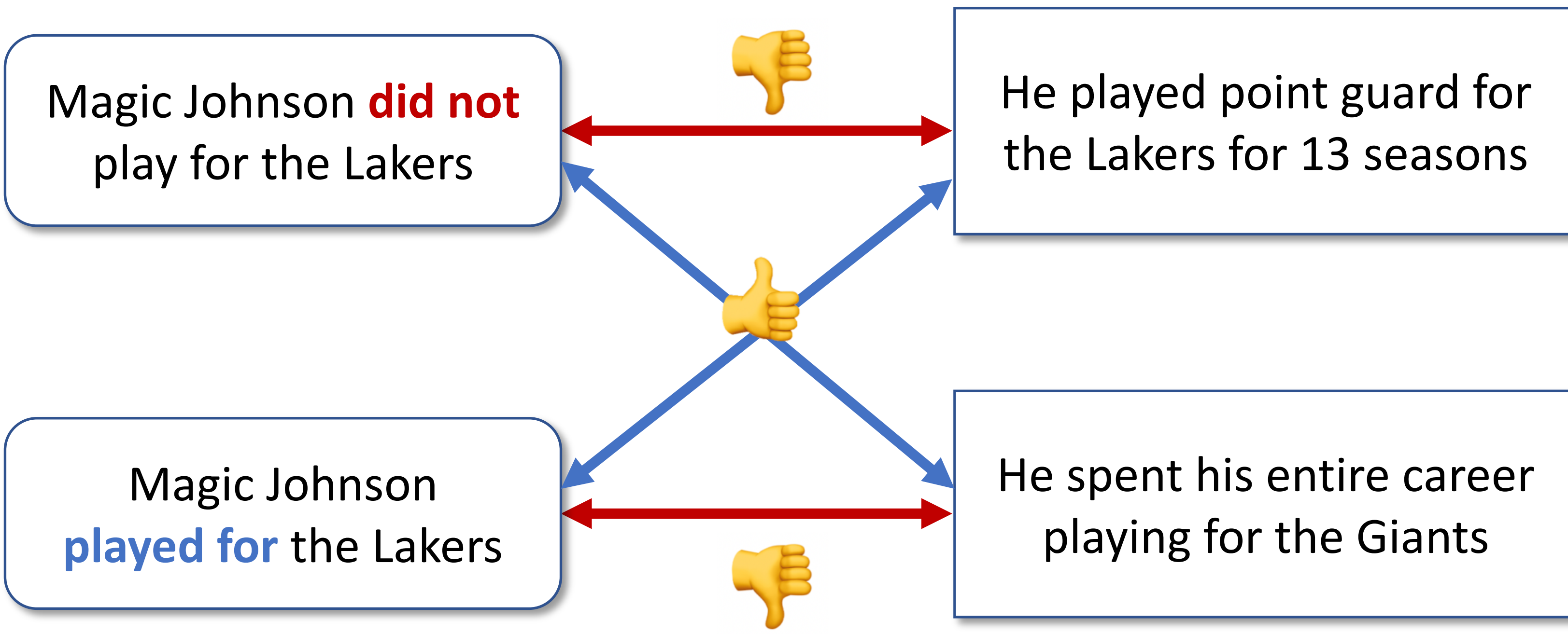
Magic Johnson **did not** play for the Lakers

He played point guard for the Lakers for 13 seasons

Creating a symmetric unbiased dataset

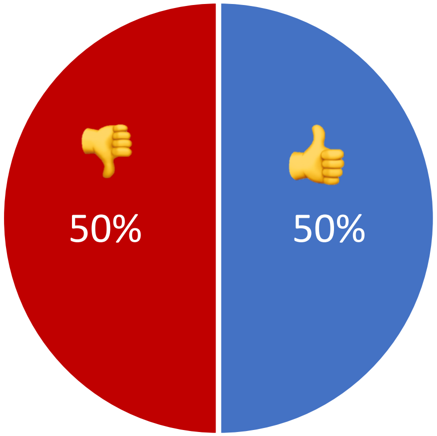


Creating a symmetric unbiased dataset



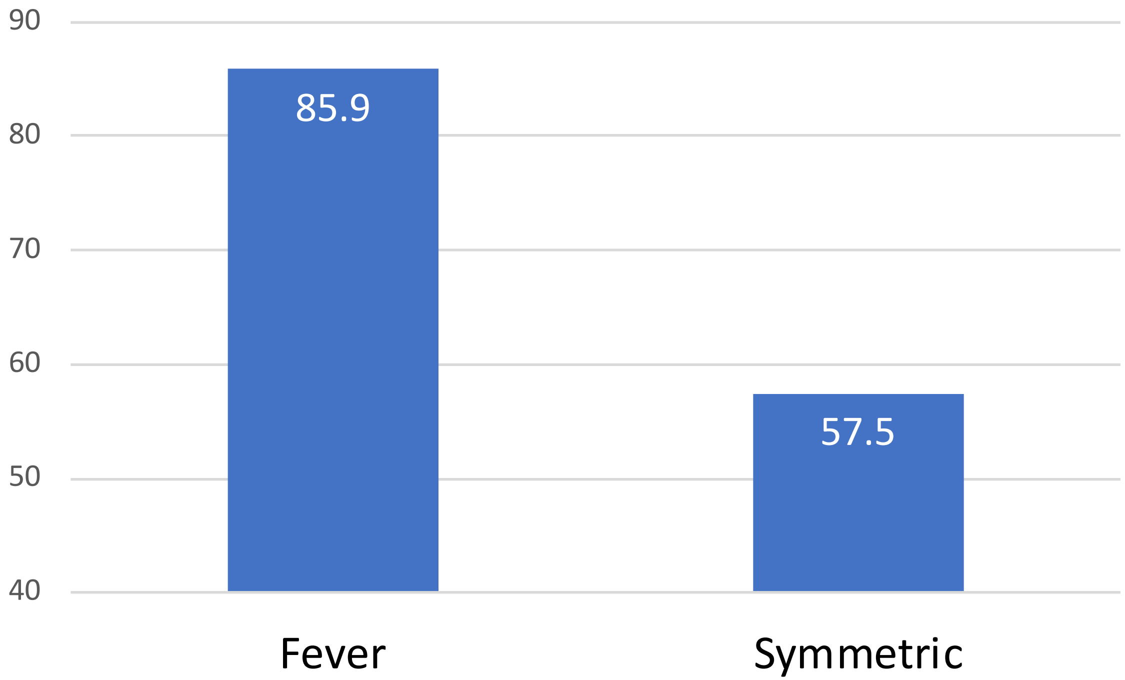
The evidence is crucial for predictions

$$p(l|w_j)$$



Performance on the symmetric dataset

Entailment results using Bert



Debiasing the training

By reweighting the training samples

Regularizing the training

- Defining the bias of phrase j towards label l:

$$b_j^l = p(l|w_j) = \frac{\sum_{i=1}^{n} \mathbb{1}_{\left[w_j^{(i)}\right]} \cdot \mathbb{1}_{\left[y^{(i)}=l\right]}}{\sum_{i=1}^{n} \mathbb{1}_{\left[w_j^{(i)}\right]}}$$

- Setting weights (α) for each training sample:

$$b_j^l = \frac{\sum_{i=1}^{n} \mathbb{1}_{\left[w_j^{(i)}\right]} \cdot \mathbb{1}_{\left[y^{(i)}=l\right]} \cdot \left(1+\alpha^{(i)}\right)}{\sum_{i=1}^{n} \mathbb{1}_{\left[w_j^{(i)}\right]} \cdot \left(1+\alpha^{(i)}\right)}$$

Regularizing the training

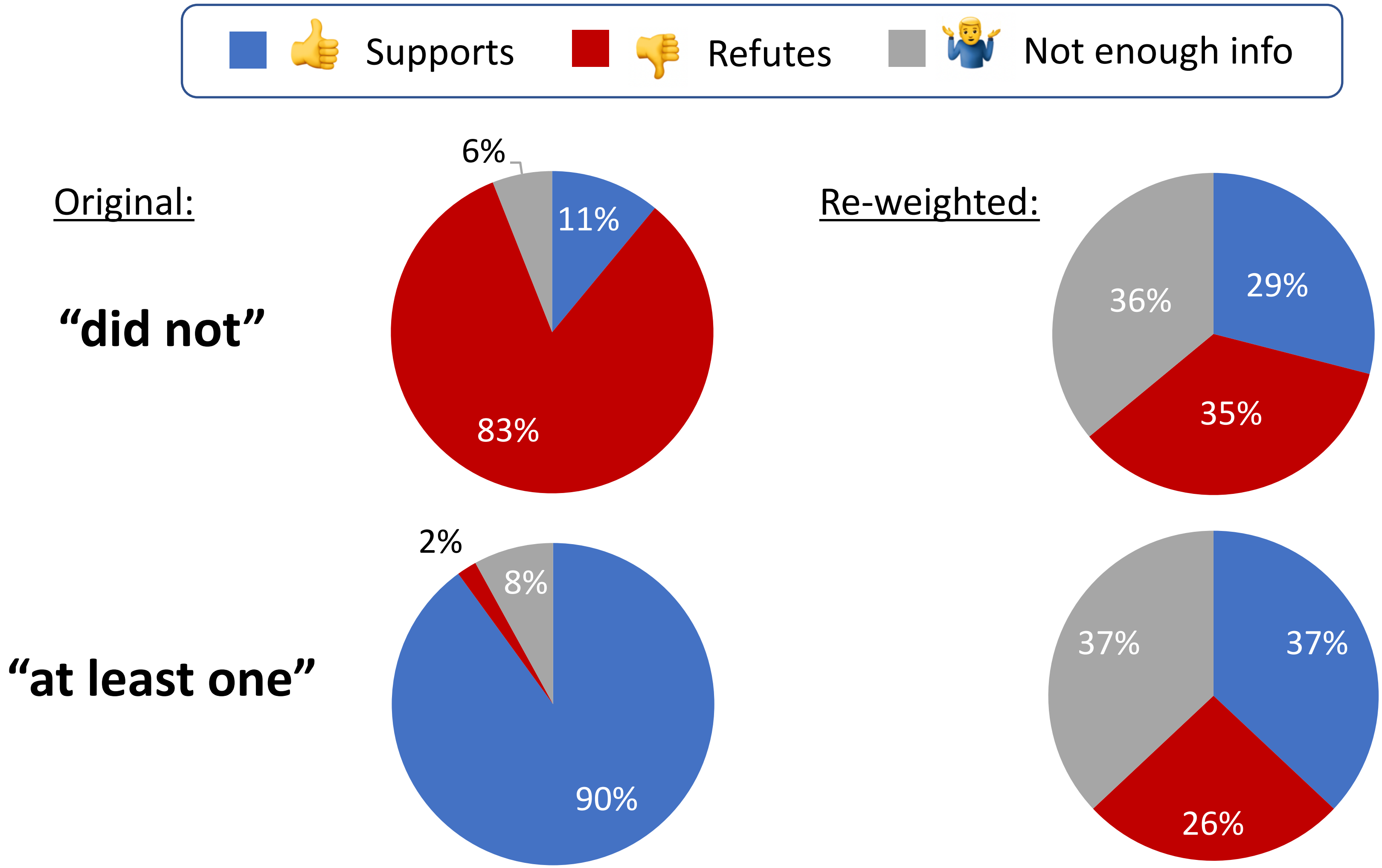
- Learning the weights by optimizing:

$$\min\left(\sum_{j=1}^{|V|} \max_{l}\left(b_j^l\right) + \lambda\|\vec{\alpha}\|_2\right)$$

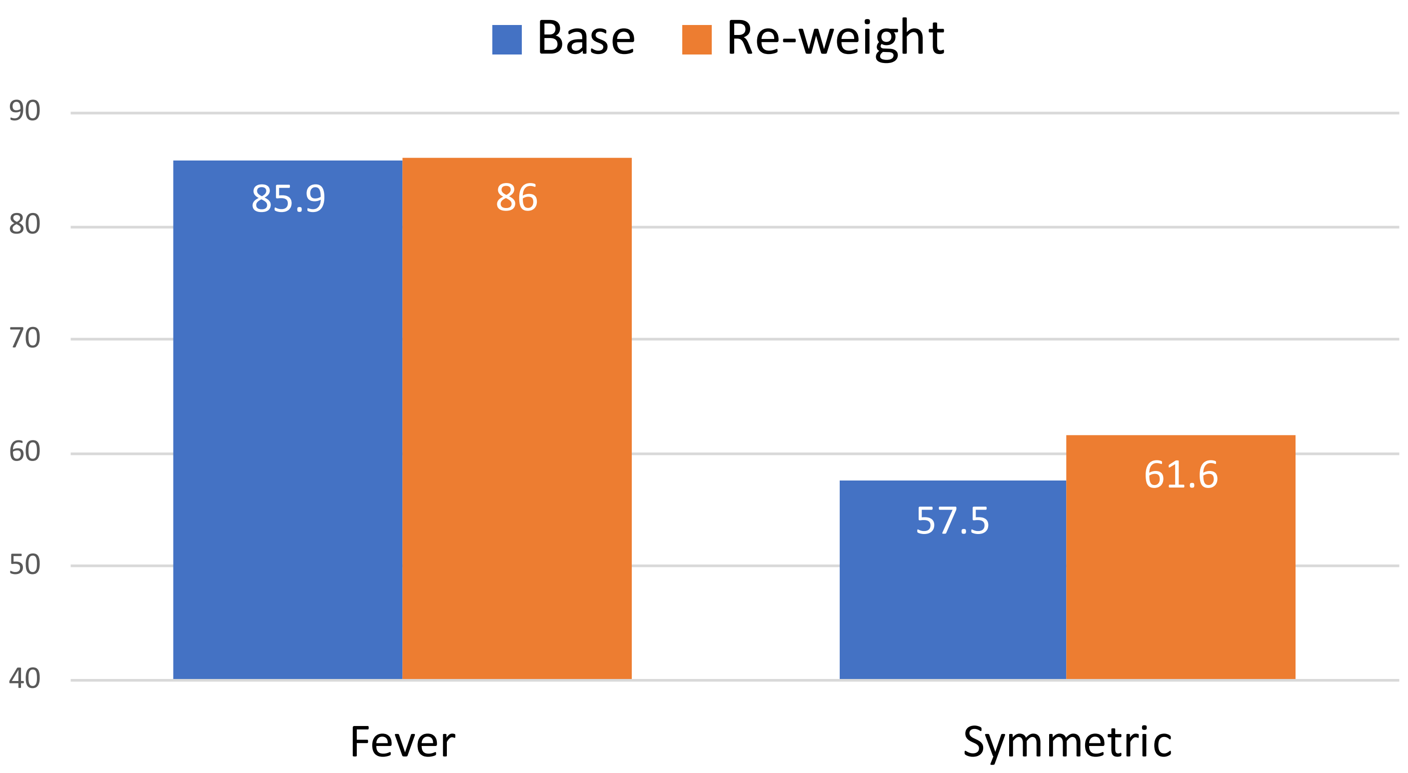
- **Re-weighted** loss function:

$$\sum_{i=1}^{n}\left(1+\alpha^{(i)}\right)\cdot\mathcal{L}\left(x^{(i)}, y^{(i)}\right)$$

Statistical cues are alleviated

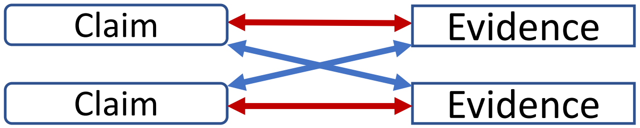


Performance on symmetric dataset



Towards Debiasing Fact Verification Models

- Bias in FEVER
  - give-away phrases in the claims
- Symmetric dataset



- Alleviating the bias
  - Re-weighting the training samples: $$\min\left(\sum_j \max_l \left(b_j^l\right)\right)$$

Code and data:
https://github.com/TalSchuster/FeverSymmetric

@TalSchuster
@darshj_shah