Model Interpretability

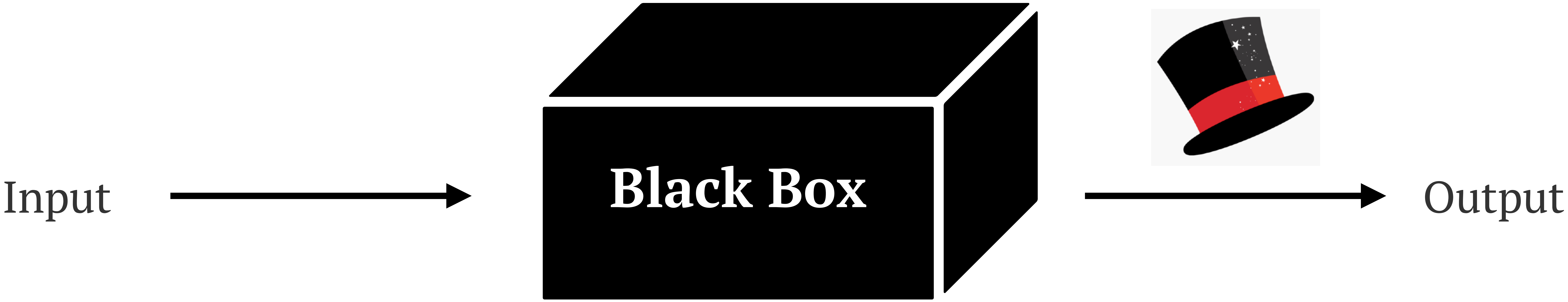
COMP3361 — Week 12

Lingpeng Kong

Department of Computer Science, The University of Hong Kong

Many materials from Zhiyong Wu with Special Thanks!

Black Box Models



Input

Black Box

Output

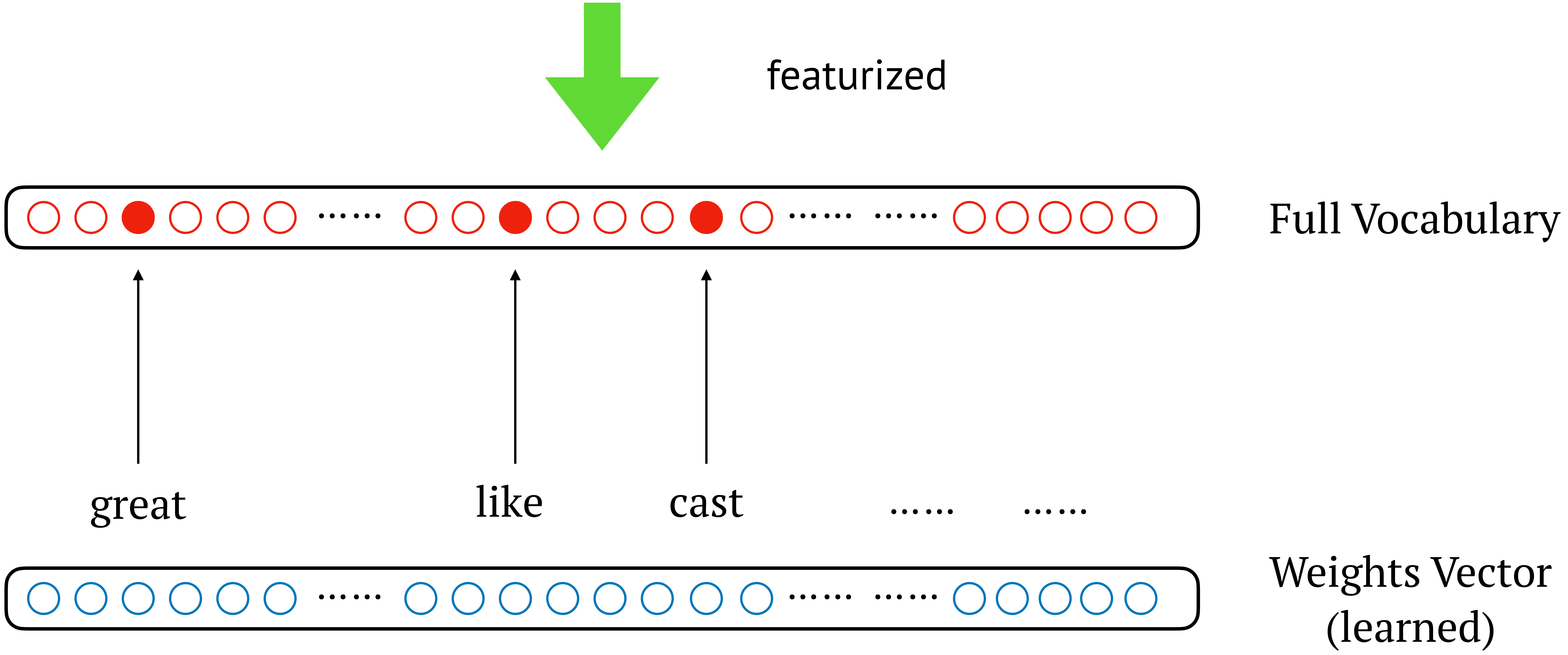
End-to-End Models

Linear Models in NLP

A Great movie ! I really like the cast .

Words:

A
great
movie
!
I
really
like
its
cast
.



featurized

Full Vocabulary

great

like

cast

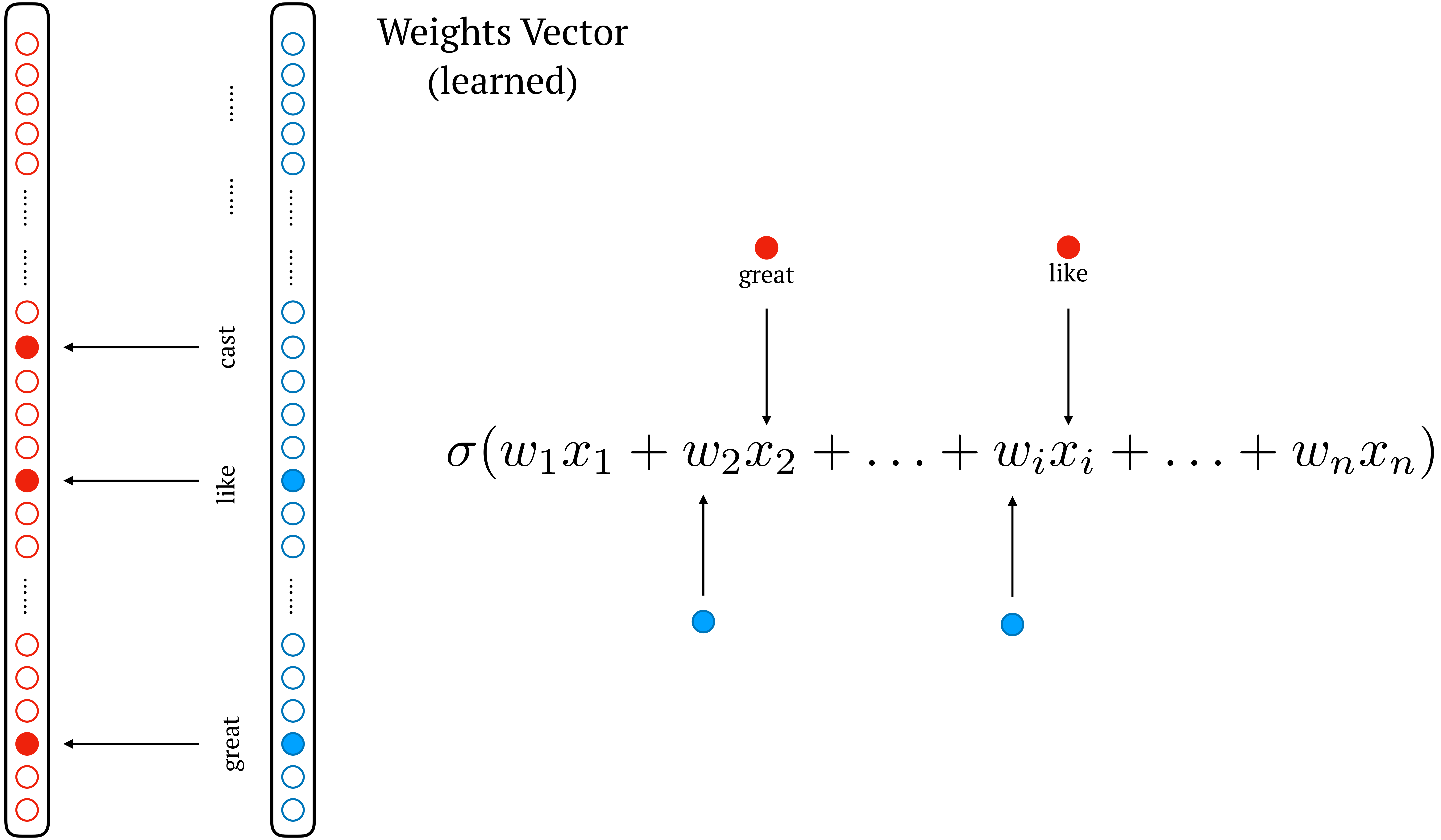
…… ……

Weights Vector
(learned)

Linear Models in NLP

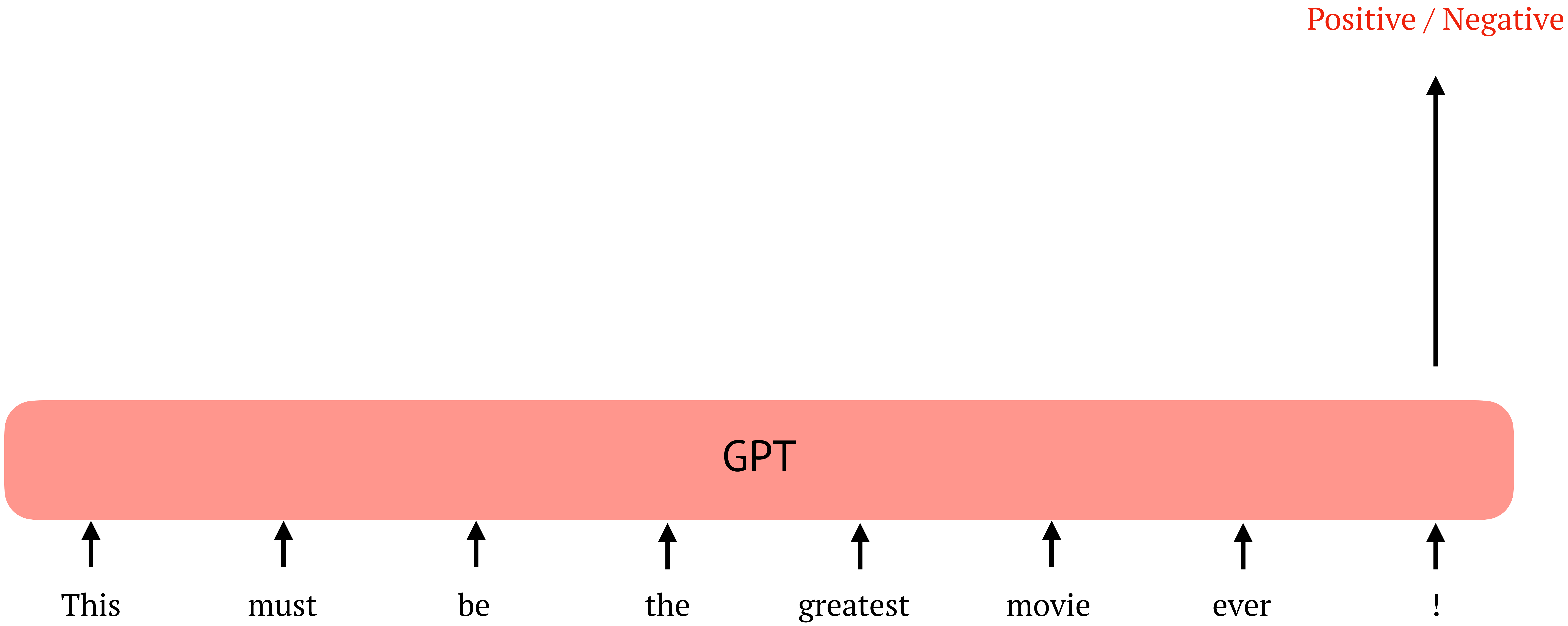
Full Vocabulary

Weights Vector (learned)

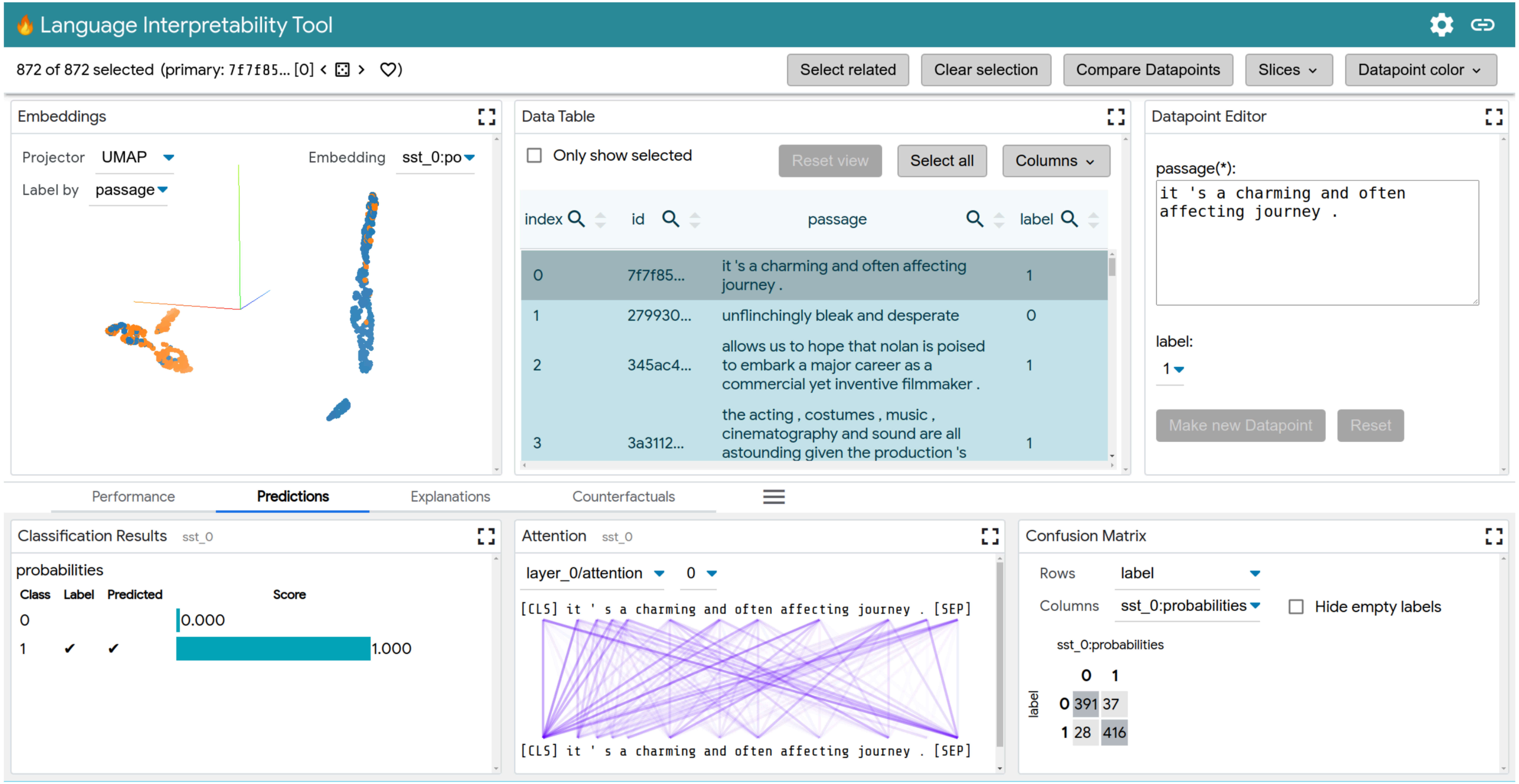


$$\sigma(w_1 x_1 + w_2 x_2 + \ldots + w_i x_i + \ldots + w_n x_n)$$

Neural Models in NLP



Neural Models in NLP



Projector: UMAP

Embedding: sst_0:po

Label by: passage

☐ Only show selected

index	id	passage	label
0	7f7f85...	it 's a charming and often affecting journey .	1
1	279930...	unflinchingly bleak and desperate	0
2	345ac4...	allows us to hope that nolan is poised to embark a major career as a commercial yet inventive filmmaker .	1
3	3a3112...	the acting , costumes , music , cinematography and sound are all astounding given the production 's	1

passage(*):

```
it 's a charming and often
affecting journey .
```

label: 1

probabilities

Class	Label	Predicted	Score
0			0.000
1	✔	✔	1.000

layer_0/attention 0

Rows: label

Columns: sst_0:probabilities

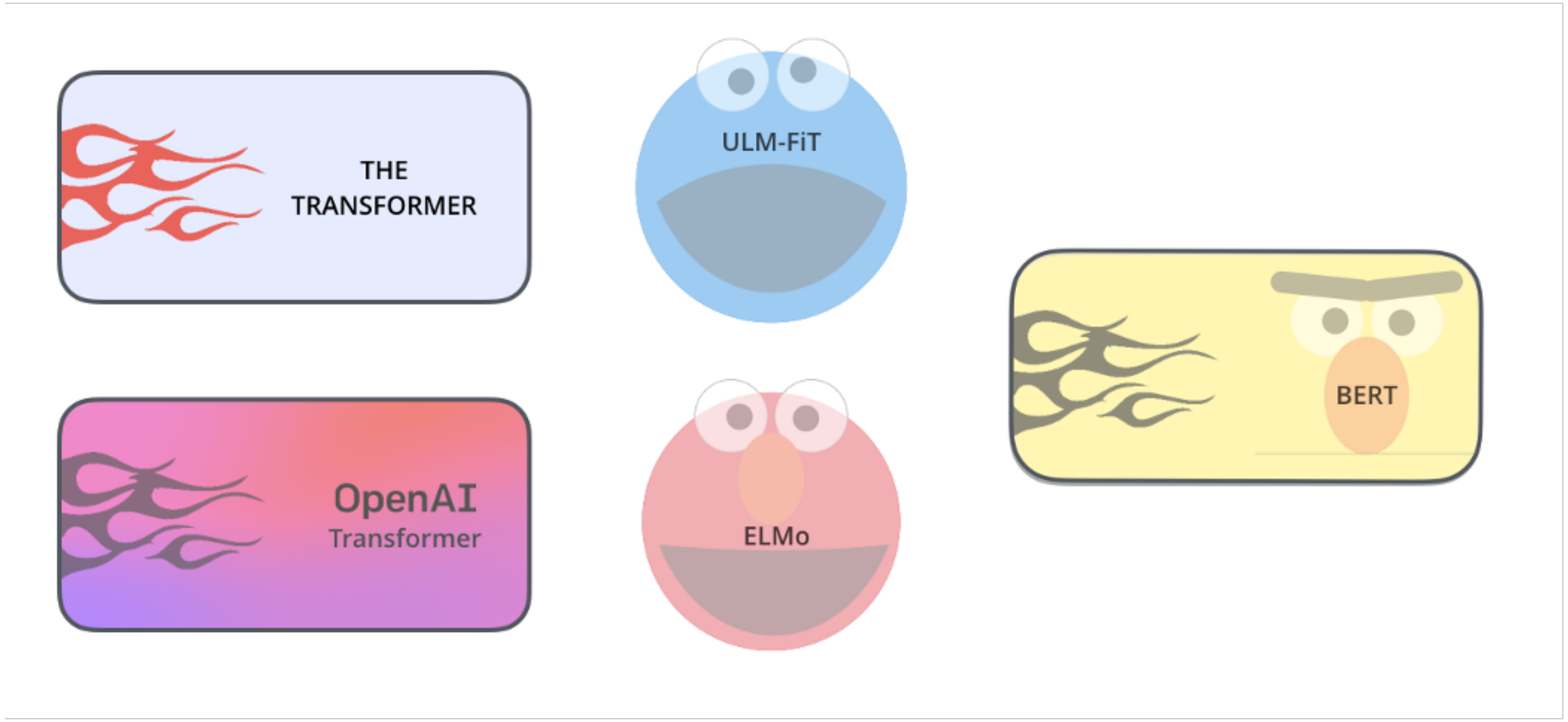
☐ Hide empty labels

sst_0:probabilities

label	0	1
0	391	37
1	28	416

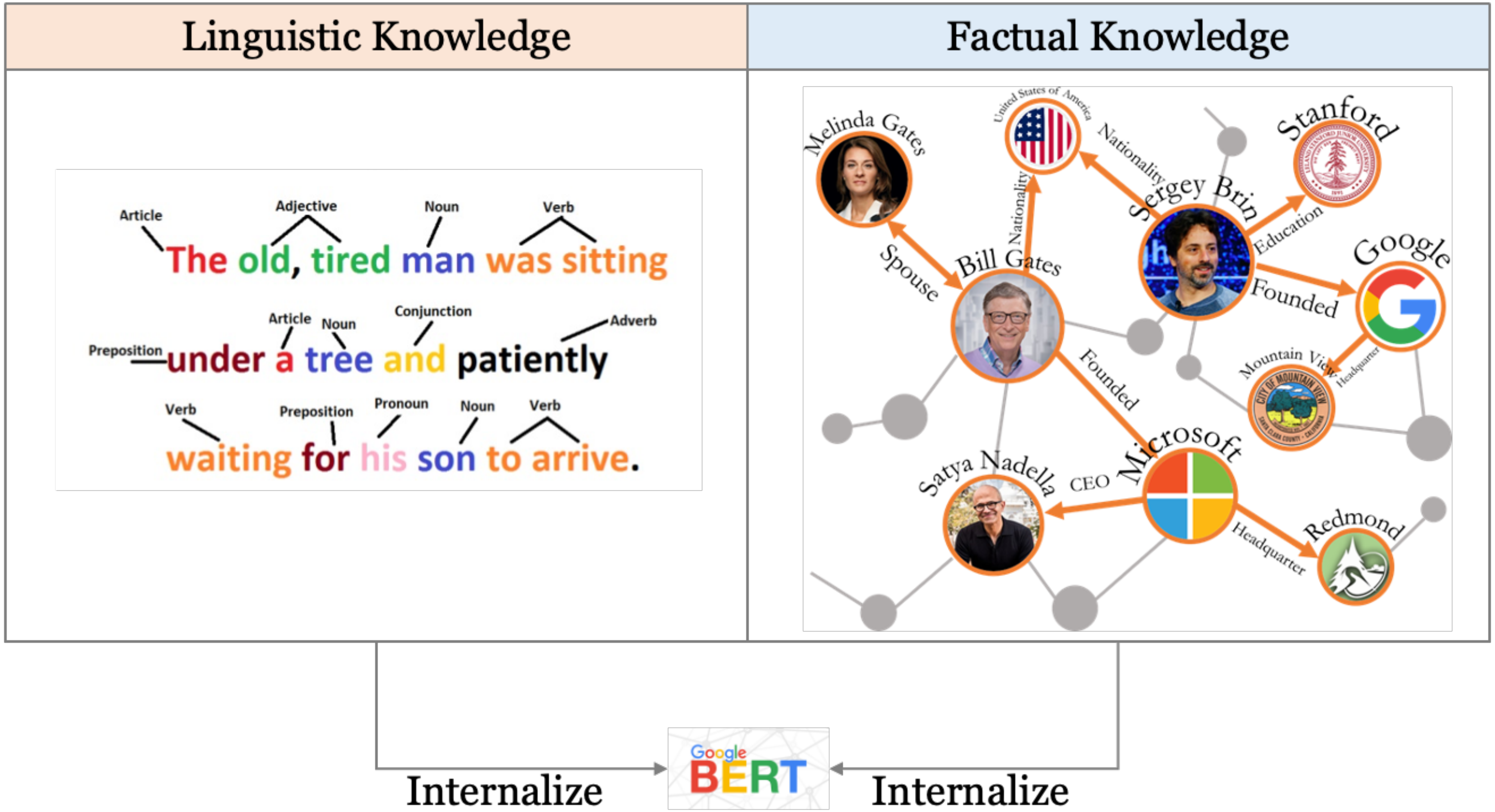
https://www.youtube.com/watch?v=j0OfBWFUqIE

The Great Sucess of Large-scale Language Models



Why language models are so powerful?

A Hypothesis



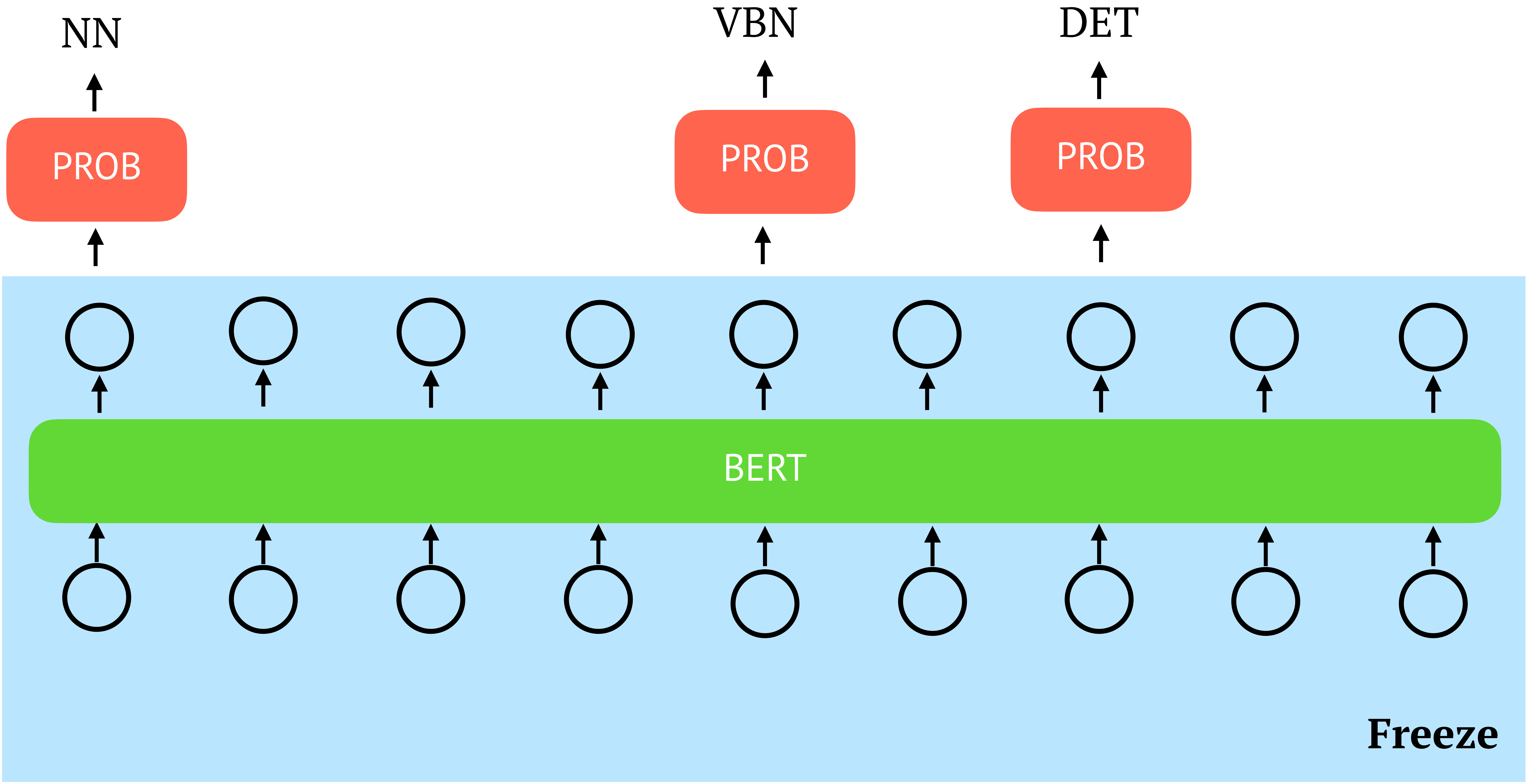
Linguistic Knowledge

Factual Knowledge

Internalize

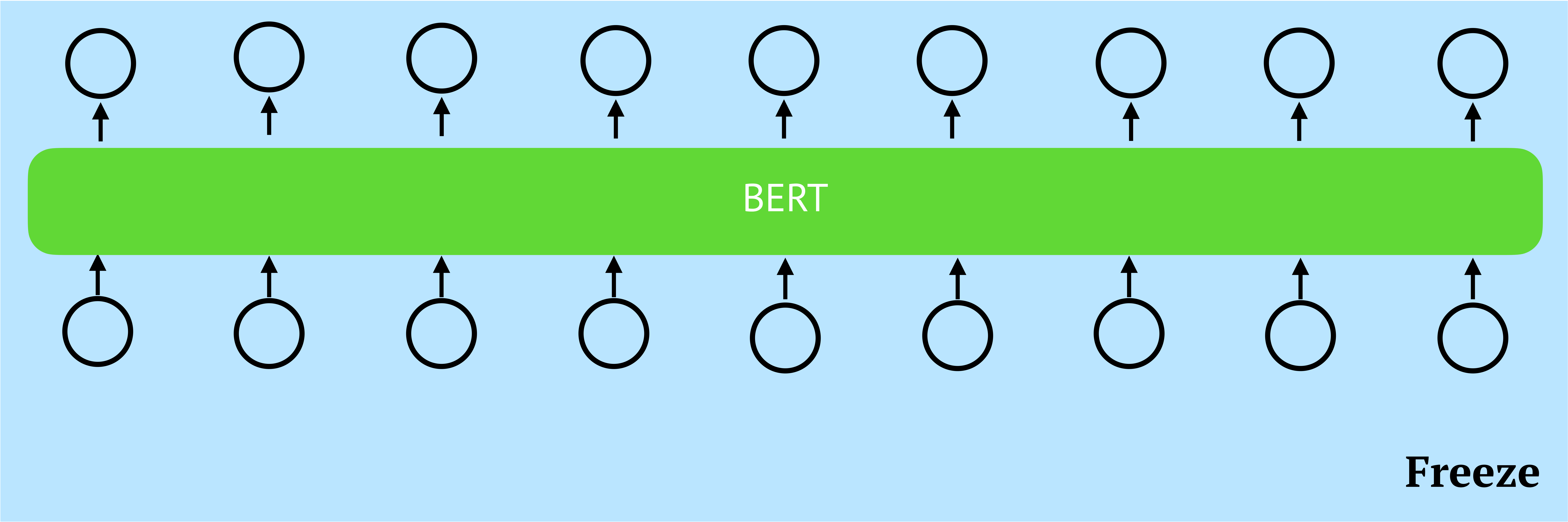
Internalize

Probing

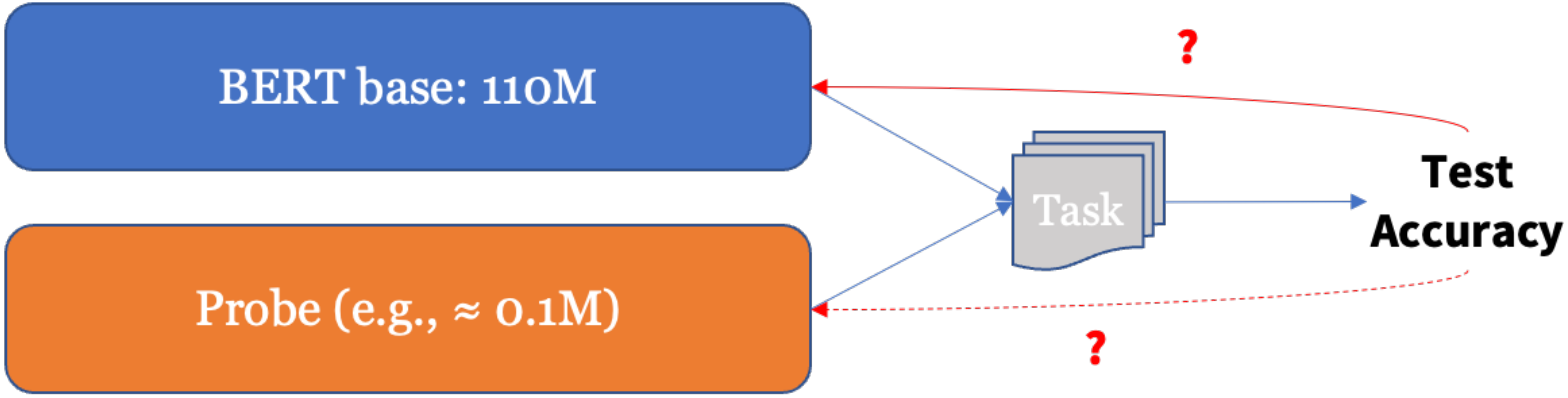


Probing

arc_score (○, ○) = PROB ○ + PROB ○ + PROB

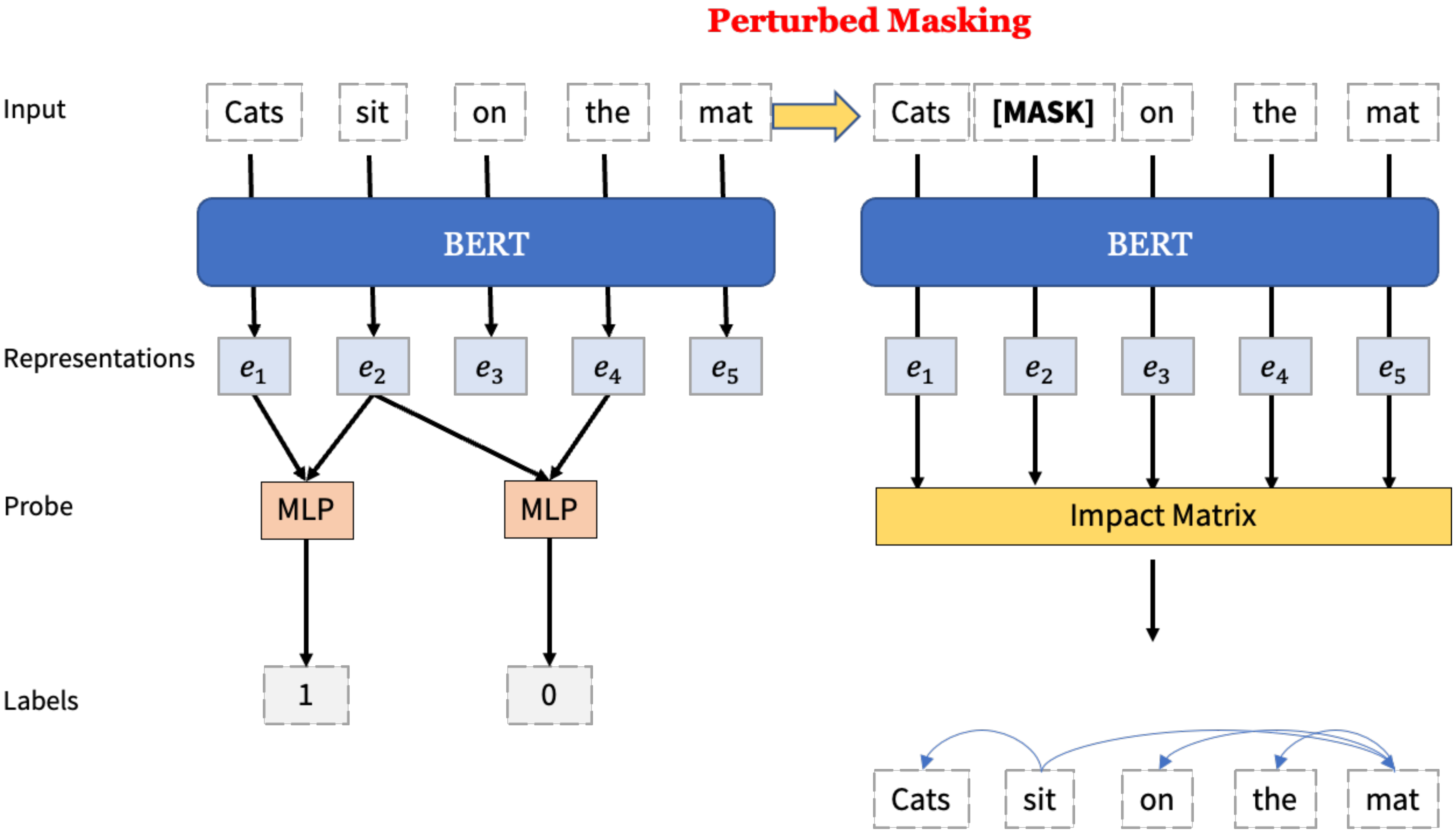


Model Probing



Shall we give credit to the **representation**? and/or the **probe**?

Parameter-free Probing



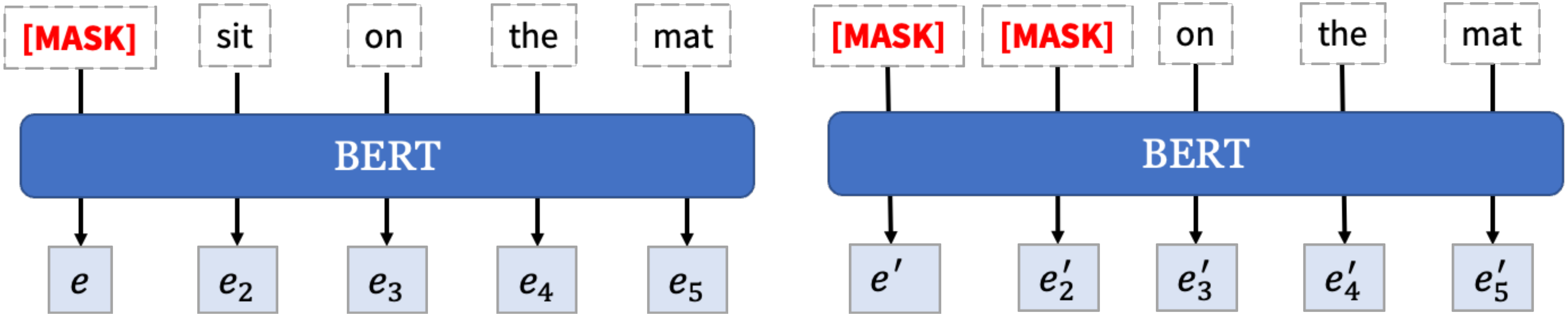
(Wu et al., 2020)

Parameter-free Probing

Example: Calculate impact *sit* has on *Cats*: $\mathrm{f}\,(Cats|sit) = d(e, e')$

$e = \mathrm{E}\,(Cats|S\backslash\{Cats\})$

$e' = \mathrm{E}\,(Cats|S\backslash\{Cats, sit\})$



(Wu et al., 2020)

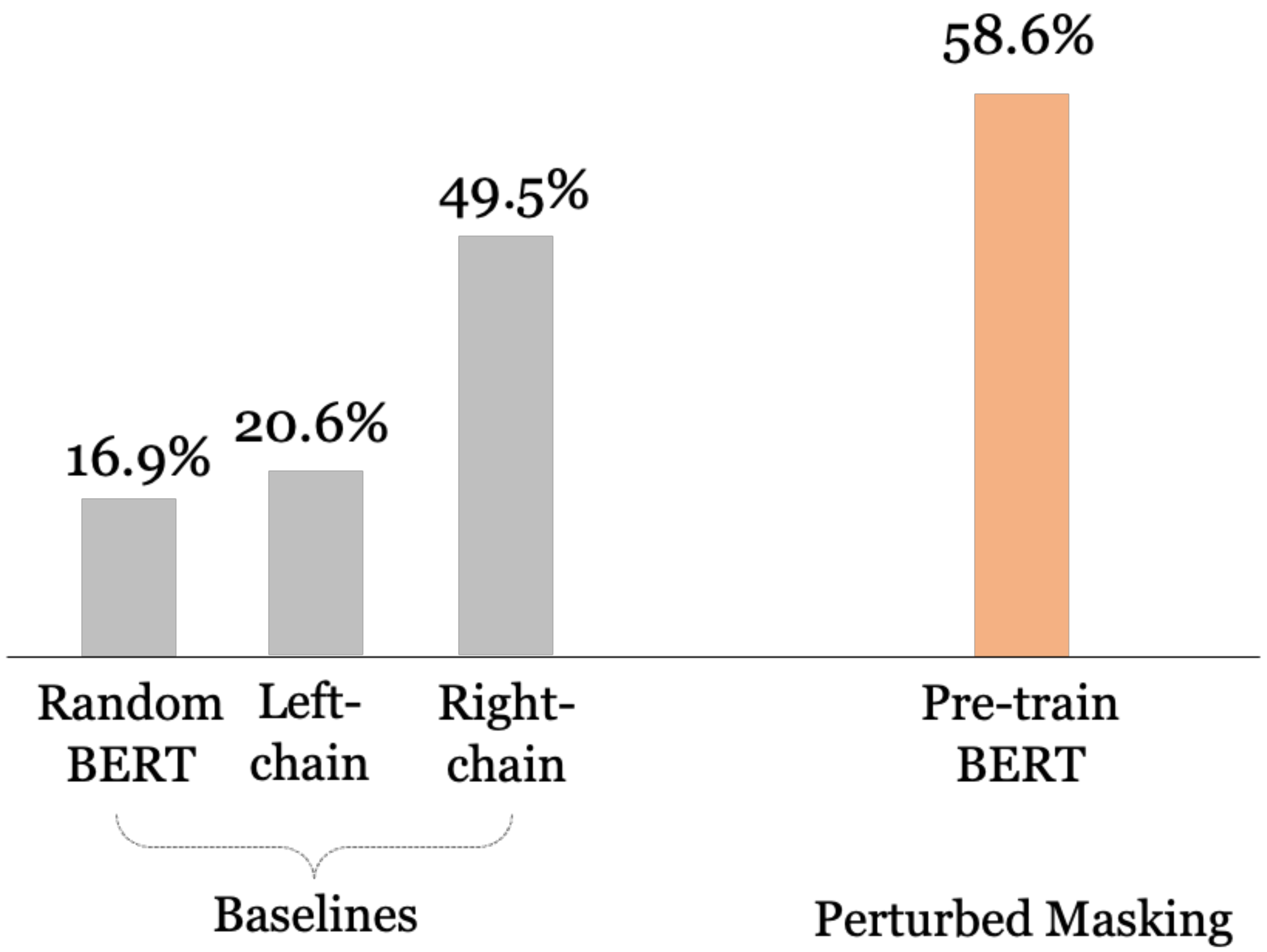
Parameter-free Probing

	Cats	sit	on	the	mat
Cats	-	f(Cat,sit)	f(Cat,on)	f(Cat,the)	f(Cat,mat)
sit	f(sit, Cats)	-	…	…	…
on	f(on, Cats)	…	-	…	…
the	f(the, Cats)	…	…	-	…
mat	f(mat, Cats)	…	…	…	-

Supervised Probe: learning to map representations to task
Here: Impact Matrix + task specific algorithm => task

(Wu et al., 2020)

Dependency Knowledge

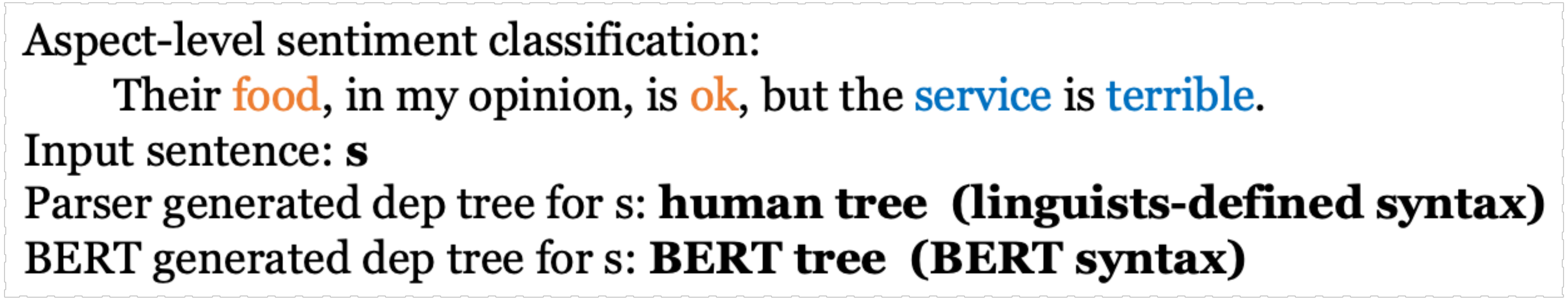


Quality of the extracted tree structure on WSJ10 (UAS score)

Would BERT learn **better** dependency structures?

(Wu et al., 2020)

Empirical Evaluation of the BERT Syntax



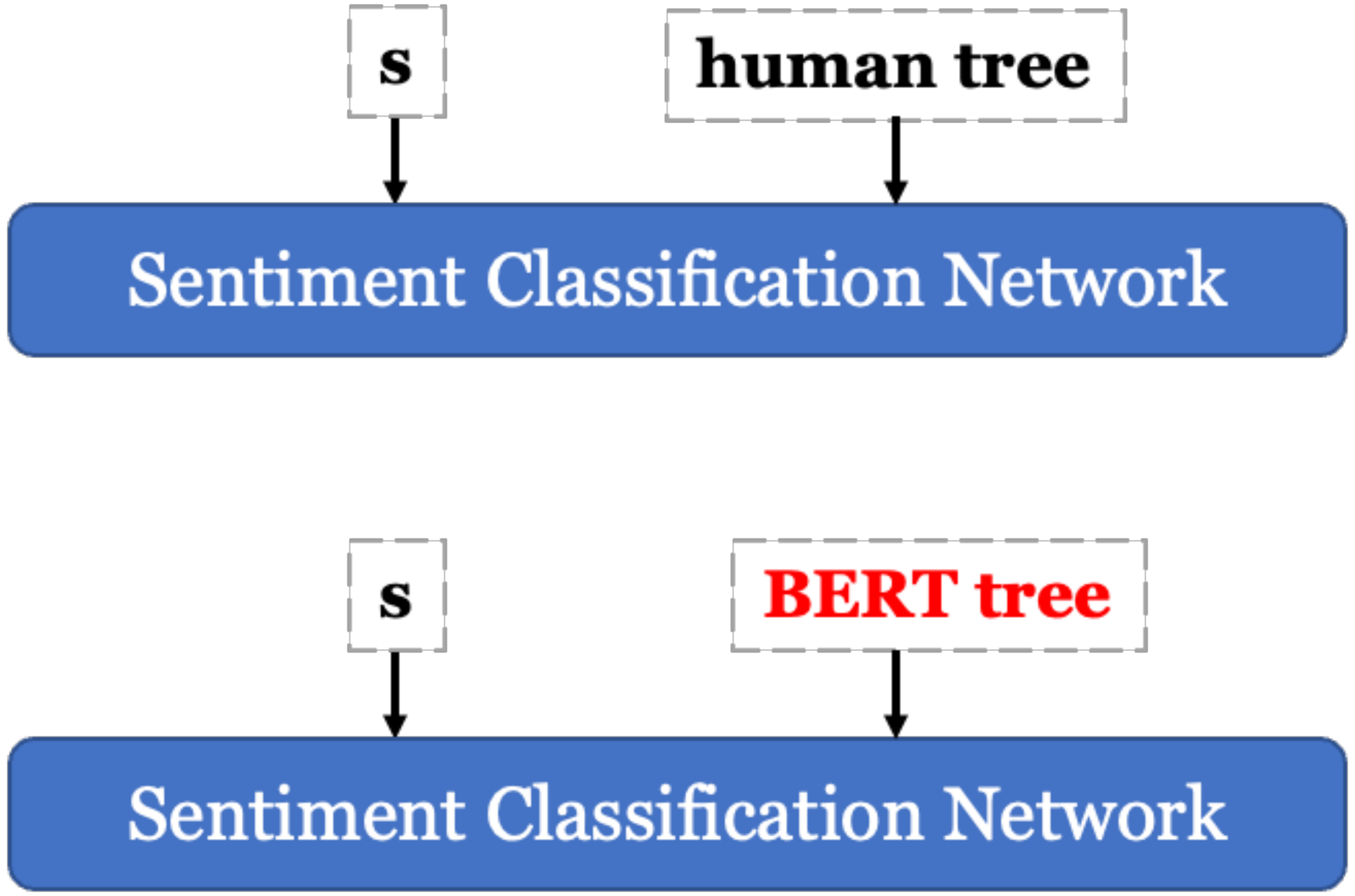
Aspect-level sentiment classification:

Their food, in my opinion, is ok, but the service is terrible.

Input sentence: **s**

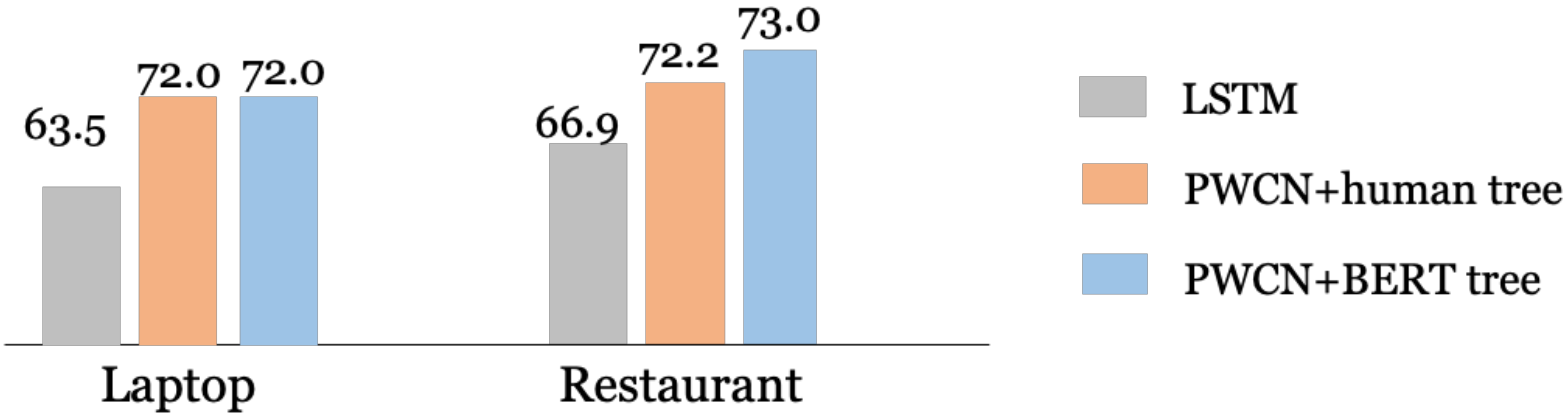
Parser generated dep tree for s: **human tree (linguists-defined syntax)**

BERT generated dep tree for s: **BERT tree (BERT syntax)**


s
human tree
Sentiment Classification Network
s
BERT tree
Sentiment Classification Network

(Wu et al., 2020)

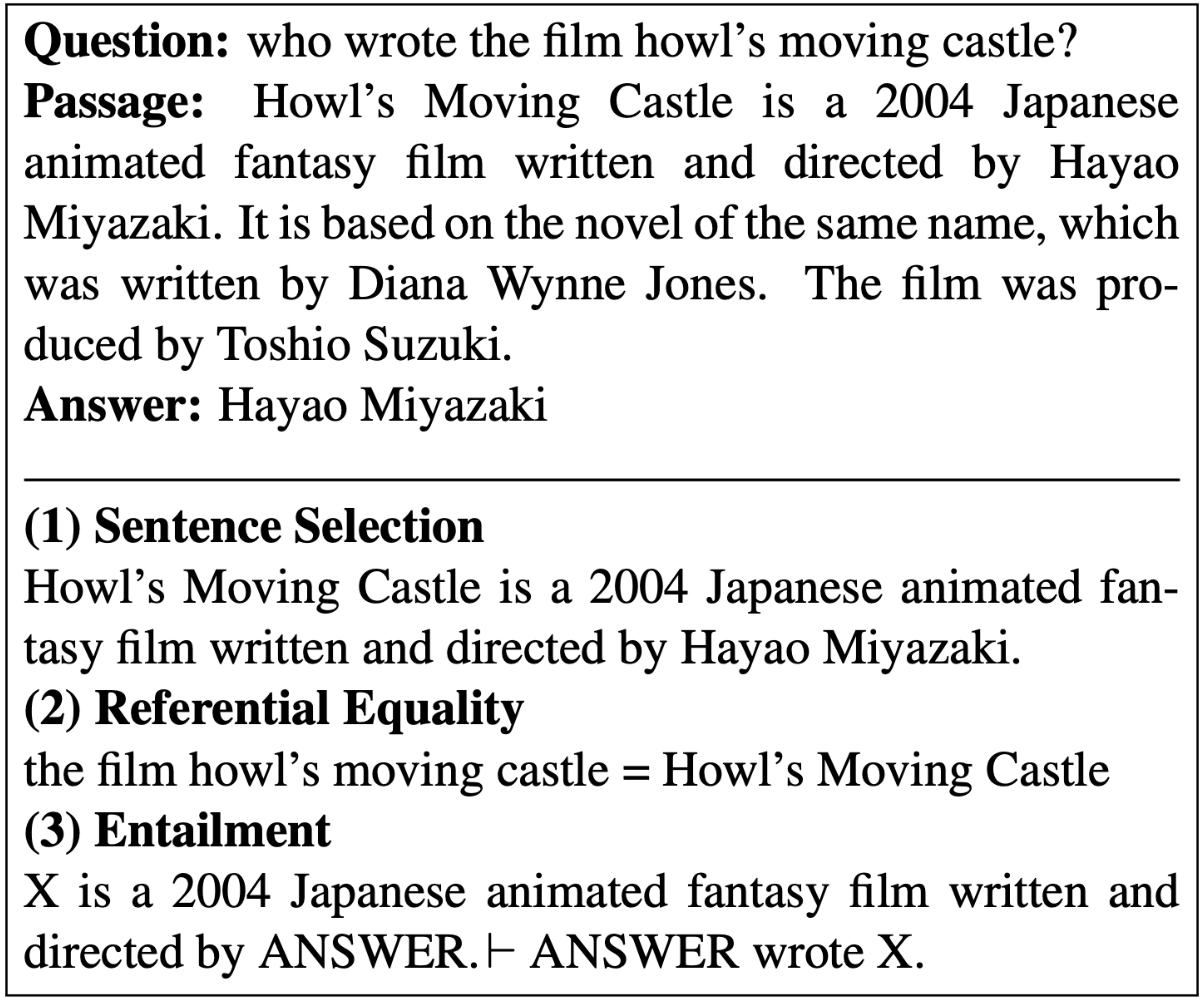
Empirical Evaluation of the BERT Syntax



Experimental Results [Marco-F1]

(Wu et al., 2020)

Towards Rationale



Question: who wrote the film howl's moving castle?
Passage: Howl's Moving Castle is a 2004 Japanese animated fantasy film written and directed by Hayao Miyazaki. It is based on the novel of the same name, which was written by Diana Wynne Jones. The film was produced by Toshio Suzuki.
Answer: Hayao Miyazaki

(1) Sentence Selection
Howl's Moving Castle is a 2004 Japanese animated fantasy film written and directed by Hayao Miyazaki.
(2) Referential Equality
the film howl's moving castle = Howl's Moving Castle
(3) Entailment
X is a 2004 Japanese animated fantasy film written and directed by ANSWER. ⊢ ANSWER wrote X.

(Lamm et al., 2020)