Probing spurious features

Evaluate Dataset

Evaluate Model

Please upload your dataset with the following format :

1. The data file should be in csv data format.
2. The data should contain four parts: ID, PREMISE, HYPOTHESIS AND LABEL.

For example:

ID	PREMISE	HYPOTHESIS	LABEL
0			Entailment

Training data file Upload

Test data file Upload

If you don't have your own dataset, you can choose a popular dataset in the following:

SNLI QNLI MNLI

Evaluate Dataset Evaluate Model

ICQ

Visualizing statistical biases in Text Inference Datasets and Models

Click to select one of the following datasets:

○ SNLI ○ QNLI ○ MNLI ○ ROC ○ COPA ○ SWAG ○ ARCT ○ ARCT_adv ○ RACE
○ RECLOR ○ CQA ○ Ubuntu ◉ Upload your own dataset

Please upload your model prediction result on test data with the following format:

1. The data file should be in txt data format.

2. The data should contain four parts: ID, PREDICTION on each line and connected with "\t".

For example: 0\tentailement\n

Test prediction file Choose File | no file selected

If you have test prediction file.

Training data file Choose File | no file selected

Test data file Choose File | no file selected

Spurious features:

- ☐ Word
- ☐ Overlap
- ☐ Typos
- ☐ Sentiment
- ☐ Negation
- ☐ Tense
- ☐ NER

Submit

ICQ

Click to select one of the following datasets:

○ SNLI ○ QNLI ○ MNLI ○ ROC ○ COPA ○ SWAG ○ ARCT ○ ARCT_adv ○ RACE
○ RECLOR ○ CQA ○ Ubuntu ◉ Upload your own dataset

Please upload your model prediction result on test data with the following format:

1. The data file should be in txt data format.

2. The data should contain four parts: ID, PREDICTION on each line and connected with "\t"

For example: 0

Click to select one of the following datasets:

○ SNLI ○ QNLI ○ MNLI ◉ ROC ○ COPA ○ SWAG ○ ARCT ○ ARCT_adv ○ RACE
○ RECLOR ○ CQA ○ Ubuntu ○ Upload your own dataset

Test predictio file

Please choose a model to test in the following:

○ Bert ○ ESIM ○ fastText

Training data file Choose File no file selected

Test data file Choose File no file selected

Spurious features:

☑ Words ☑ Overlap ☑ Typos ☑ Sentiment ☑ Negation ☑ Tense ☑ NER

Submit

ICQ

Click to select one of the following datasets:

○ SNLI ○ QNLI ○ MNLI ◉ ROC ○ COPA ○ SWAG ○ ARCT ○ ARCT_adv ○ RACE
○ RECLOR ○ CQA ○ Ubuntu ○ Upload your own dataset

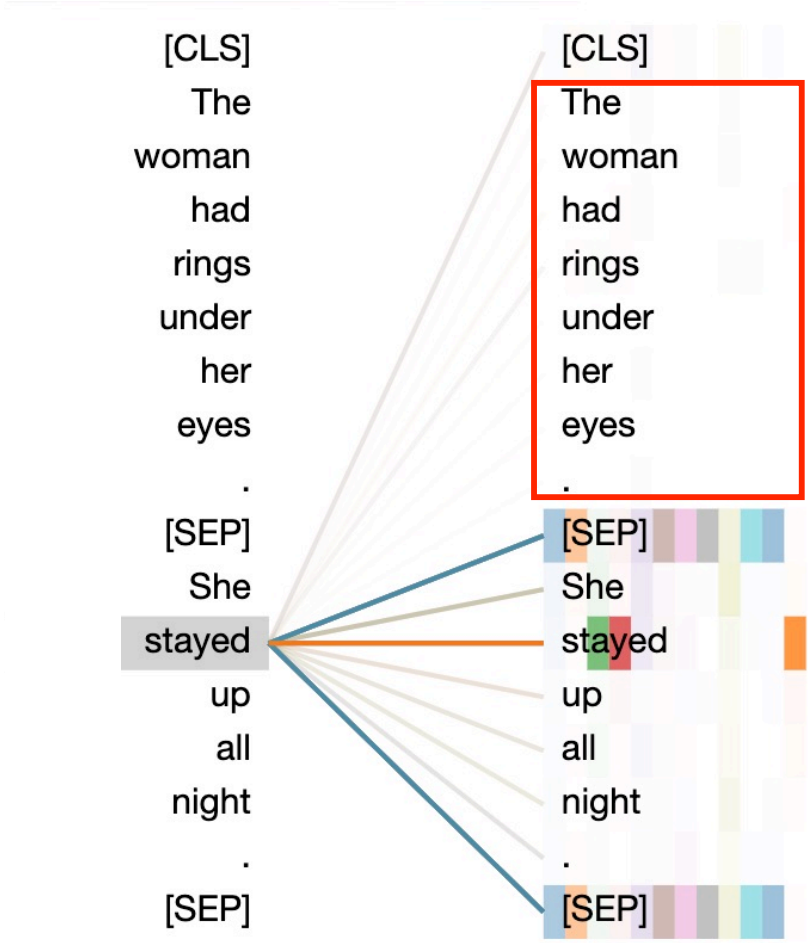
Please choose a model to test in the following:

○ Bert ○ ESIM ○ fastText

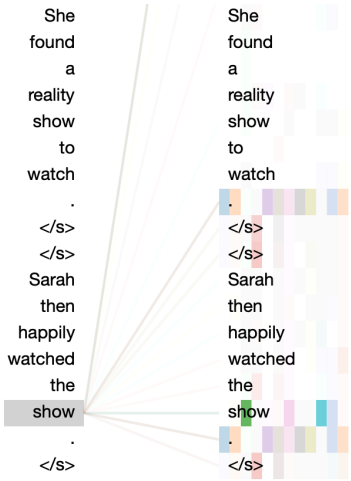
Spurious features:

☑ Words ☑ Overlap ☑ Typos ☑ Sentiment ☑ Negation ☑ Tense ☑ NER

Submit

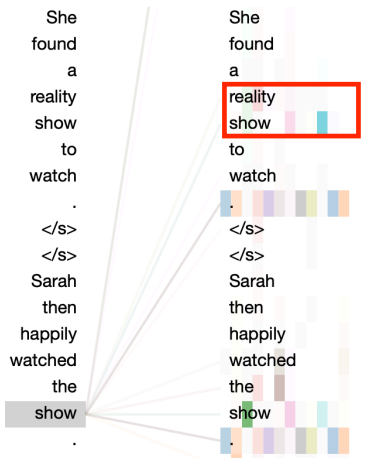

[CLS]
The
woman
had
rings
under
her
eyes
.
[SEP]
She
stayed
up
all
night
.
[SEP]
[CLS]
The
woman
had
rings
under
her
eyes
.
[SEP]
She
stayed
up
all
night
.
[SEP]

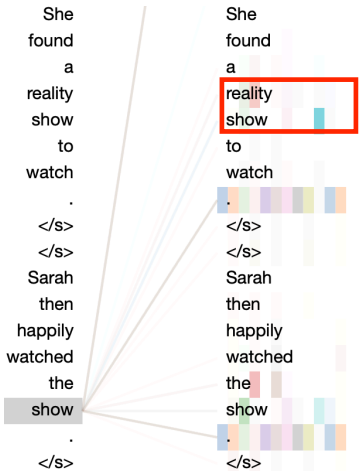
Roc bert

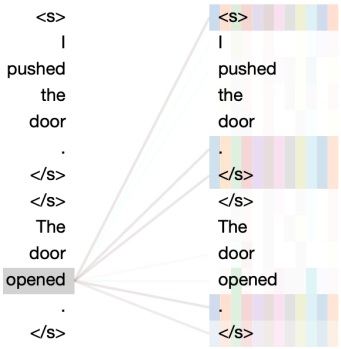

She
found
a
reality
show
to
watch
.
</s>
</s>
Sarah
then
happily
watched
the
show
.
</s>
She
found
a
reality
show
to
watch
.
</s>
</s>
Sarah
then
happily
watched
the
show
.
</s>

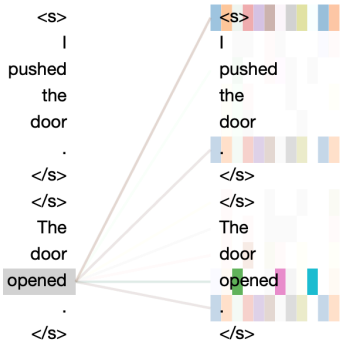

She
found
a
reality
show
to
watch
.
</s>
</s>
Sarah
then
happily
watched
the
show
.
</s>
She
found
a
reality
show
to
watch
.
</s>
</s>
Sarah
then
happily
watched
the
show
.
</s>

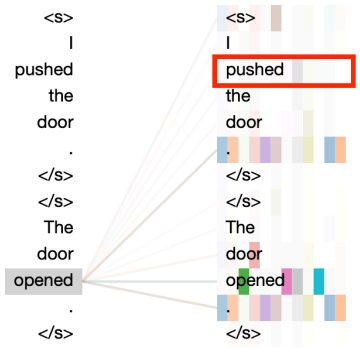

She
found
a
reality
show
to
watch
.
</s>
</s>
Sarah
then
happily
watched
the
show
.
</s>
She
found
a
reality
show
to
watch
.
</s>
</s>
Sarah
then
happily
watched
the
show
.
</s>


She
found
a
reality
show
to
watch
.
</s>
</s>
Sarah
then
happily
watched
the
show
.
She
found
a
reality
show
to
watch
.
</s>
</s>
Sarah
then
happily
watched
the
show
.

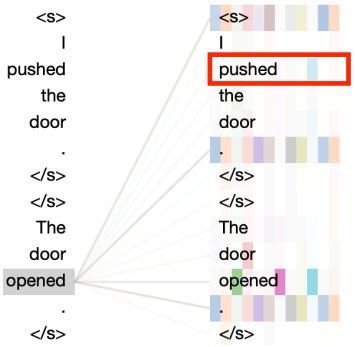

She
found
a
reality
show
to
watch
.
</s>
</s>
Sarah
then
happily
watched
the
show
.
</s>
She
found
a
reality
show
to
watch
.
</s>
</s>
Sarah
then
happily
watched
the
show
.
</s>


<s>
I
pushed
the
door
.
</s>
</s>
The
door
opened
.
</s>
<s>
I
pushed
the
door
.
</s>
</s>
The
door
opened
.
</s>

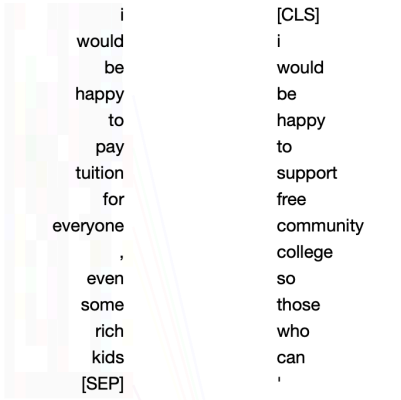

<s>
I
pushed
the
door
.
</s>
</s>
The
door
opened
.
</s>
<s>
I
pushed
the
door
.
</s>
</s>
The
door
opened
.
</s>

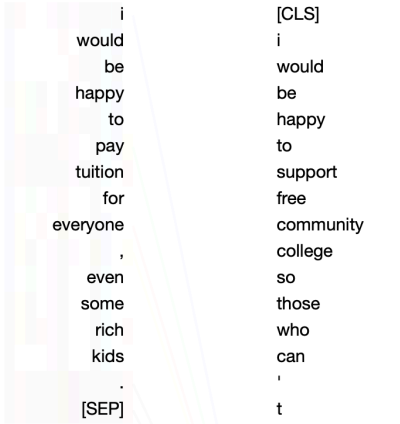

<s>
I
pushed
the
door
.
</s>
</s>
The
door
opened
.
</s>
<s>
I
pushed
the
door
.
</s>
</s>
The
door
opened
.
</s>

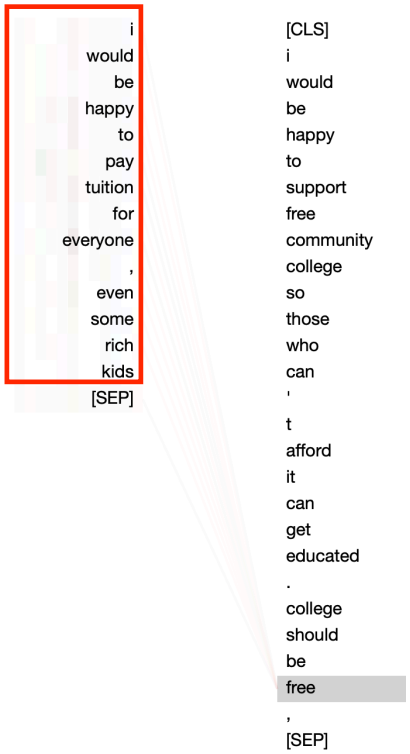

<s>
I
pushed
the
door
.
</s>
</s>
The
door
opened
.
</s>
<s>
I
pushed
the
door
.
</s>
</s>
The
door
opened
.
</s>


<s>
I
pushed
the
door
.
</s>
</s>
The
door
opened
.
</s>
<s>
I
pushed
the
door
.
</s>
</s>
The
door
opened
.
</s>

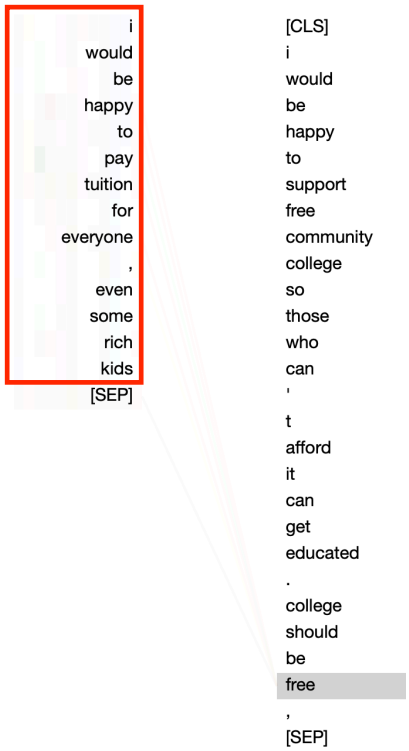
ARCT-bert


i
would
be
happy
to
pay
tuition
for
everyone
,
even
some
rich
kids
[SEP]
[CLS]
i
would
be
happy
to
support
free
community
college
so
those
who
can
'
t
afford
it
can
get
educated
.
college
should
be
free
,
[SEP]

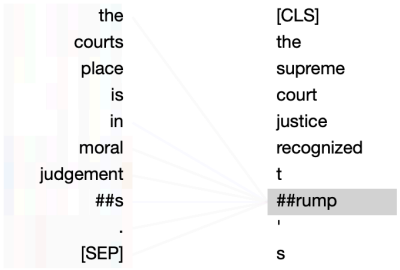

i
would
be
happy
to
pay
tuition
for
everyone
,
even
some
rich
kids
.
[SEP]
[CLS]
i
would
be
happy
to
support
free
community
college
so
those
who
can
'
t
afford
it
can
get
educated
.
college
should
be
free
,
[SEP]

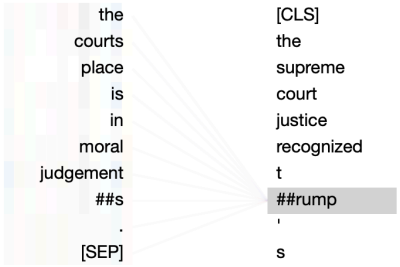

i
would
be
happy
to
pay
tuition
for
everyone
,
even
some
rich
kids
[SEP]
[CLS]
i
would
be
happy
to
support
free
community
college
so
those
who
can
'
t
afford
it
can
get
educated
.
college
should
be
free
,
[SEP]

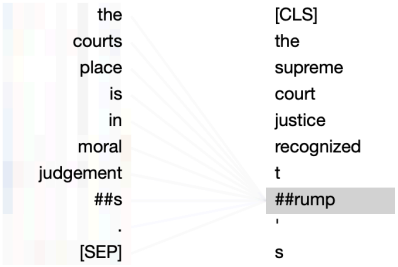

i
would
be
happy
to
pay
tuition
for
everyone
,
even
some
rich
kids
[SEP]
[CLS]
i
would
be
happy
to
support
free
community
college
so
those
who
can
'
t
afford
it
can
get
educated
.
college
should
be
free
,
[SEP]

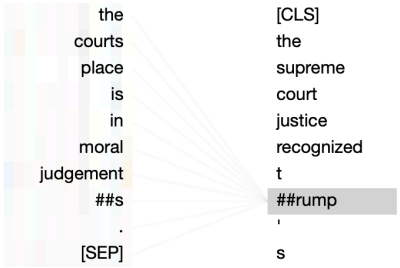

i
would
be
happy
to
pay
tuition
for
everyone
,
even
some
rich
kids
[SEP]
[CLS]
i
would
be
happy
to
support
free
community
college
so
those
who
can
'
t
afford
it
can
get
educated
.
college
should
be
free
,
[SEP]

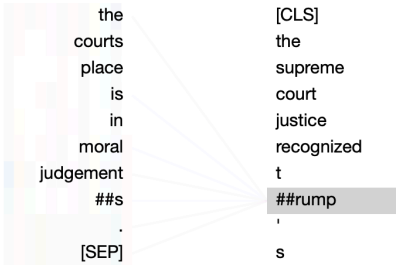
Arct- bert


the
courts
place
is
in
moral
judgement
##s
.
[SEP]
[CLS]
the
supreme
court
justice
recognized
t
##rump
'
s


the
courts
place
is
in
moral
judgement
##s
.
[SEP]
[CLS]
the
supreme
court
justice
recognized
t
##rump
'
s

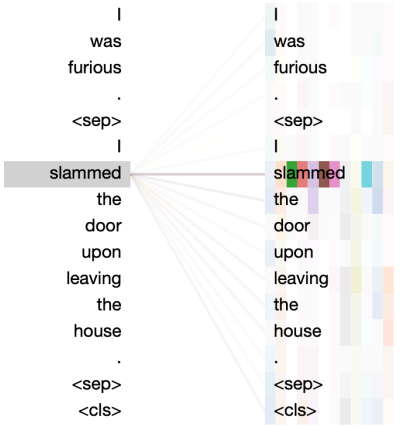

the
courts
place
is
in
moral
judgement
##s
.
[SEP]
[CLS]
the
supreme
court
justice
recognized
t
##rump
'
s

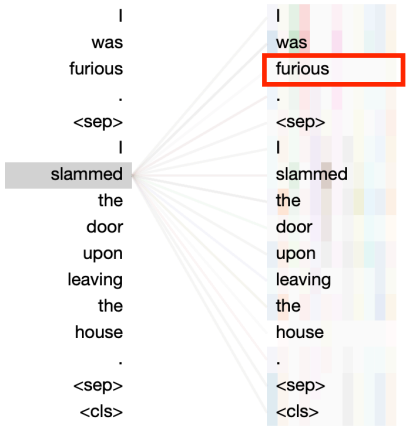

the
courts
place
is
in
moral
judgement
##s
.
[SEP]
[CLS]
the
supreme
court
justice
recognized
t
##rump
'
s

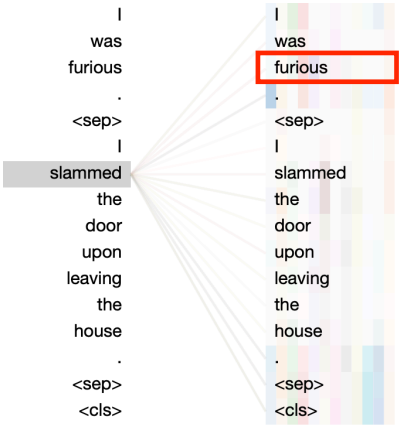

the
courts
place
is
in
moral
judgement
##s
.
[SEP]
[CLS]
the
supreme
court
justice
recognized
t
##rump
'
s

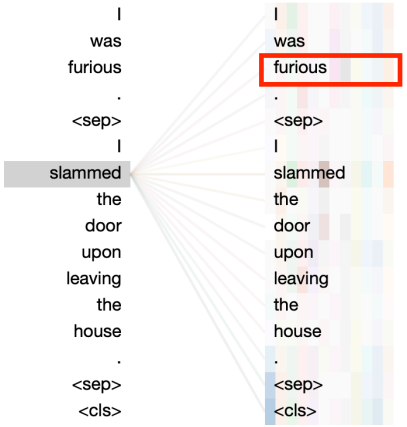
Reclor xlnet

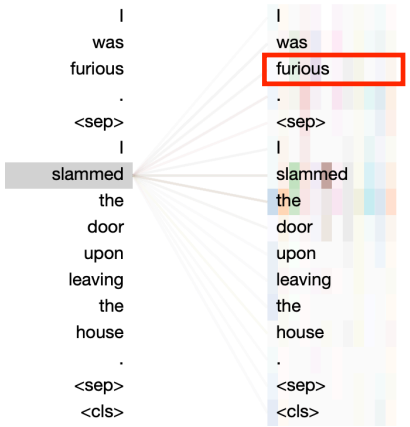
copa xlnet


I
was
furious
.
<sep>
I
slammed
the
door
upon
leaving
the
house
.
<sep>
<cls>
I
was
furious
.
<sep>
I
slammed
the
door
upon
leaving
the
house
.
<sep>
<cls>

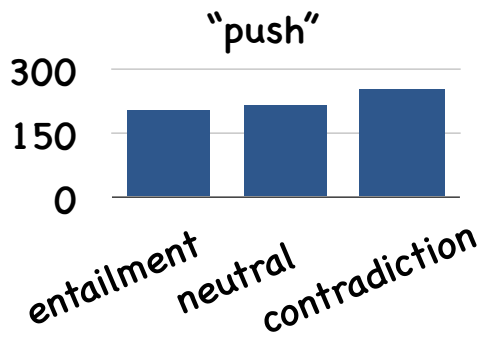

I
was
furious
.
<sep>
I
slammed
the
door
upon
leaving
the
house
.
<sep>
<cls>
I
was
furious
.
<sep>
I
slammed
the
door
upon
leaving
the
house
.
<sep>
<cls>

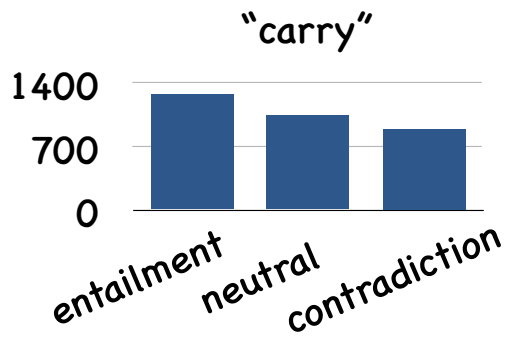

I
was
furious
.
<sep>
I
slammed
the
door
upon
leaving
the
house
.
<sep>
<cls>
I
was
furious
.
<sep>
I
slammed
the
door
upon
leaving
the
house
.
<sep>
<cls>


I
was
furious
.
<sep>
I
slammed
the
door
upon
leaving
the
house
.
<sep>
<cls>
I
was
furious
.
<sep>
I
slammed
the
door
upon
leaving
the
house
.
<sep>
<cls>


I
was
furious
.
<sep>
I
slammed
the
door
upon
leaving
the
house
.
<sep>
<cls>
I
was
furious
.
<sep>
I
slammed
the
door
upon
leaving
the
house
.
<sep>
<cls>

Features



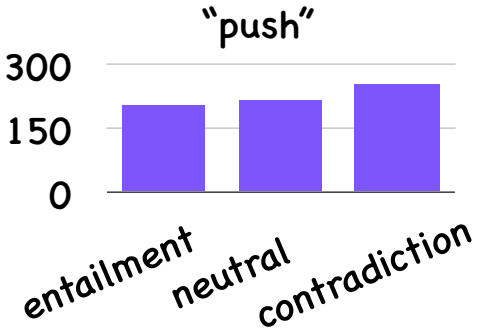


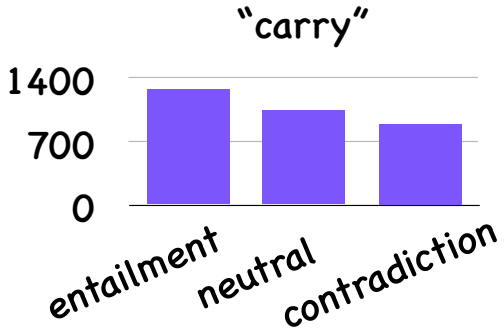
Premise: A man pulling items on a cart.

Hypothesis: A man is pushing a baby carriage. ✓

Hypothesis': A man is carrying a baby carriage. ✗

Features





Premise: A man pulling items on a cart.

Hypothesis: A man is pushing a baby carriage. ✓

Hypothesis': A man is carrying a baby carriage. ✗

Cause: The woman hummed to herself.

What was the cause for this?

Option1: She was in a good mood. ✓

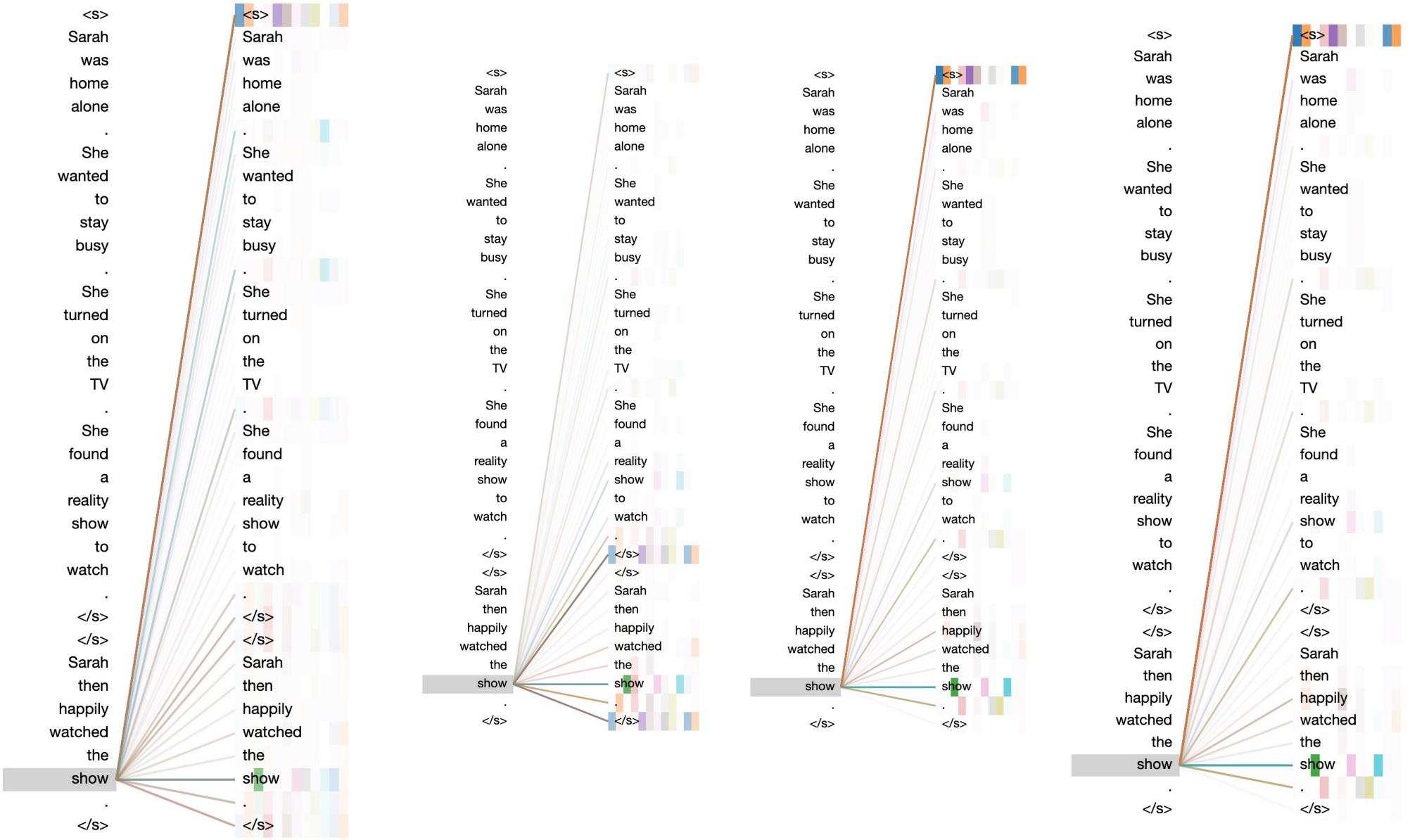
Option2: She was nervous. ✗

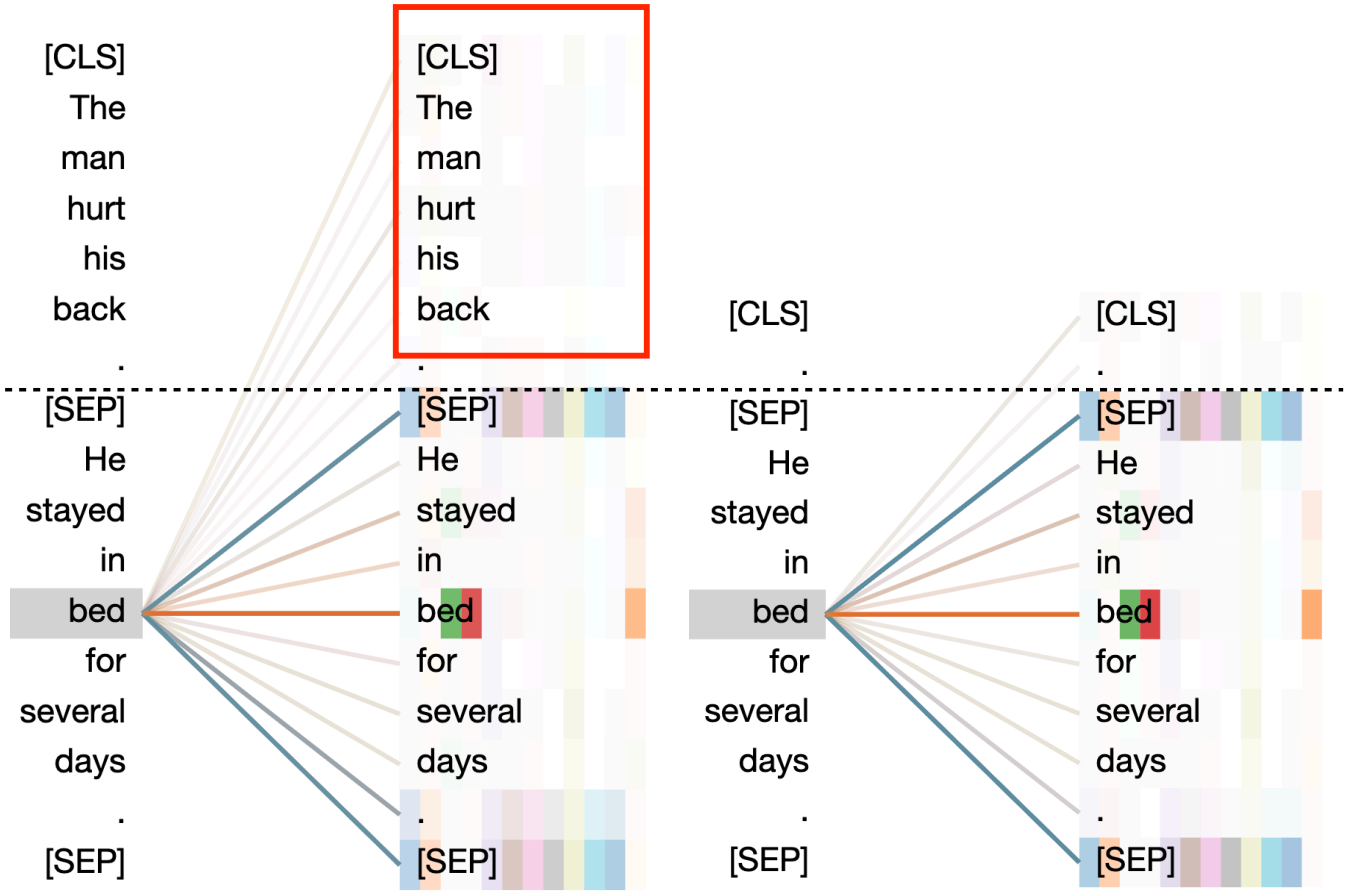
Cause: The woman trembled.

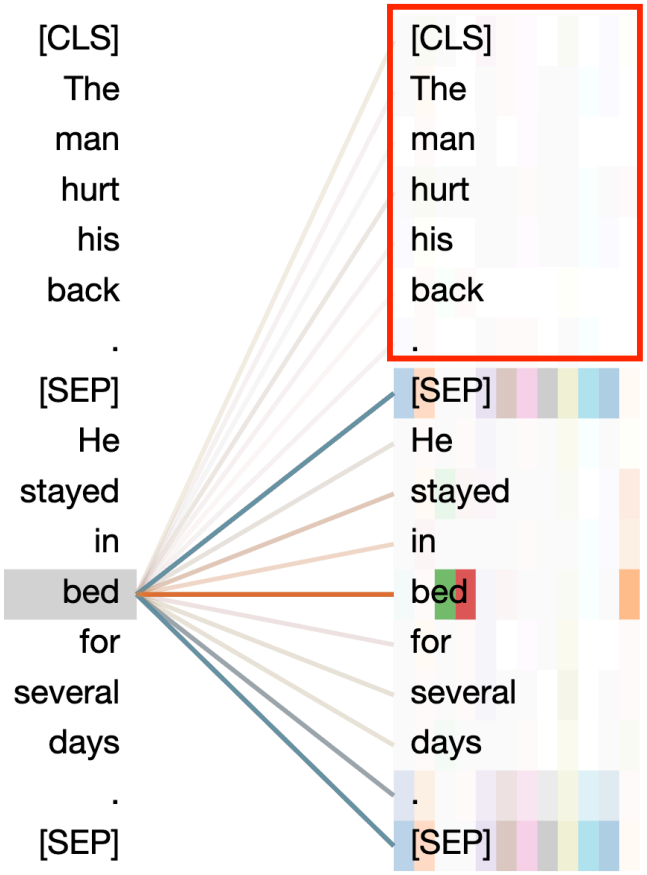
What was the cause for this?

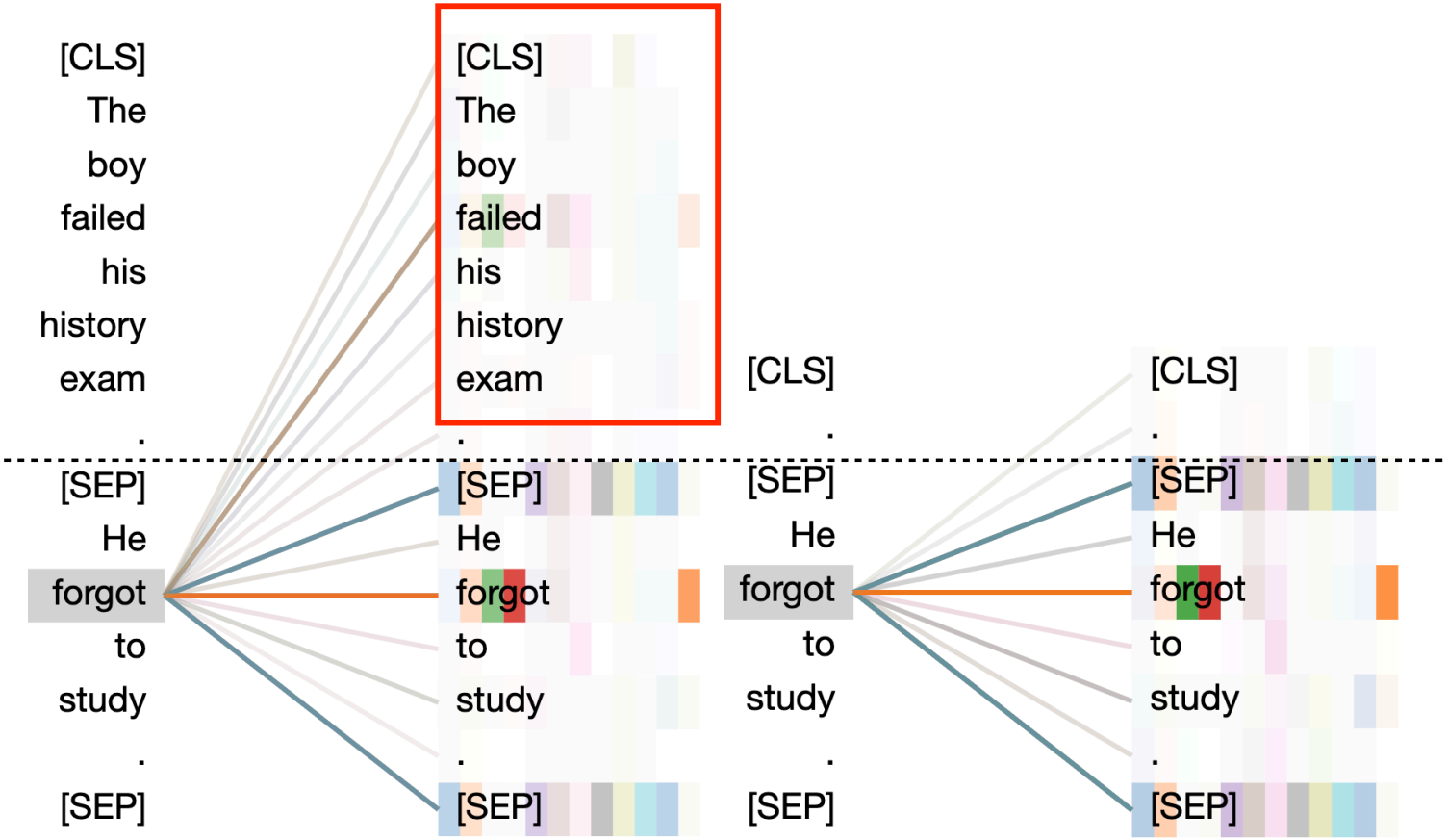
Option1: She was in a good mood. ✗

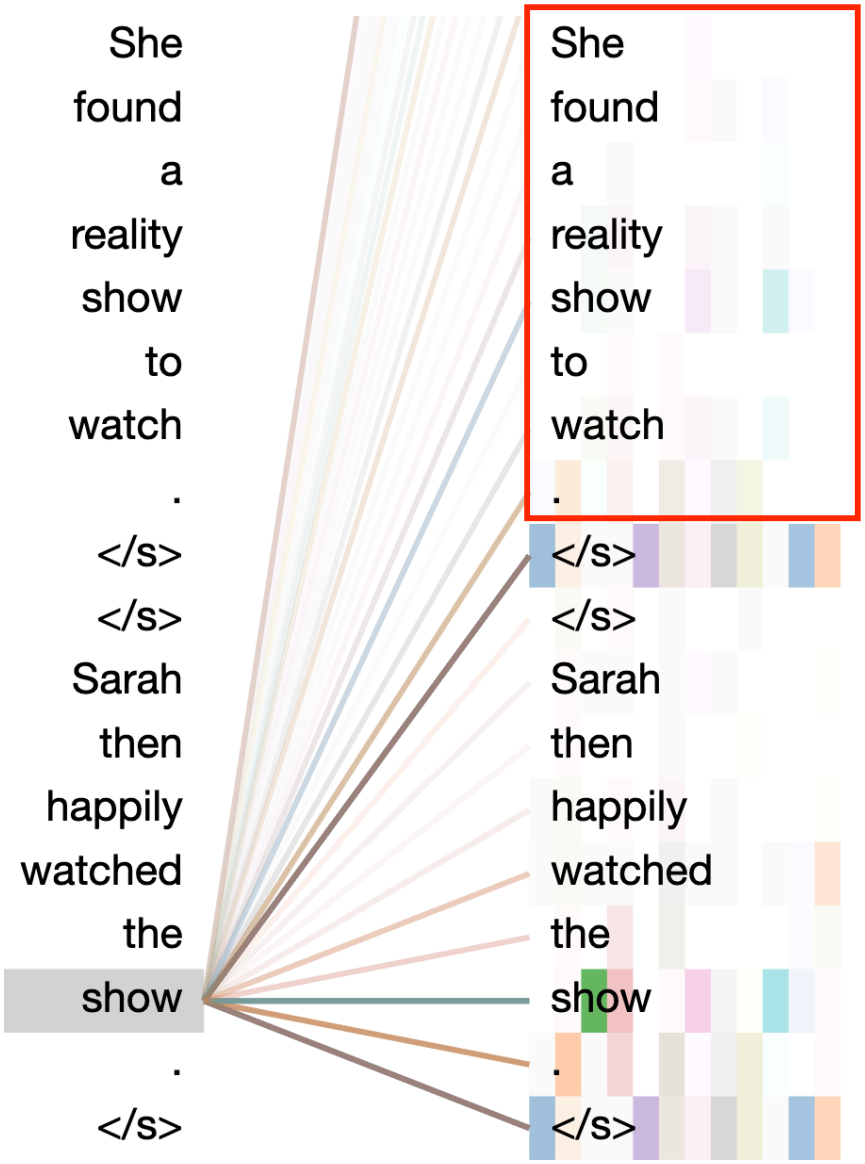
Option2: She was nervous. ✓


<s>
Sarah
was
home
alone
.
She
wanted
to
stay
busy
.
She
turned
on
the
TV
.
She
found
a
reality
show
to
watch
.
</s>
</s>
Sarah
then
happily
watched
the
show
.
</s>
<s>
Sarah
was
home
alone
.
She
wanted
to
stay
busy
.
She
turned
on
the
TV
.
She
found
a
reality
show
to
watch
.
</s>
</s>
Sarah
then
happily
watched
the
show
.
</s>

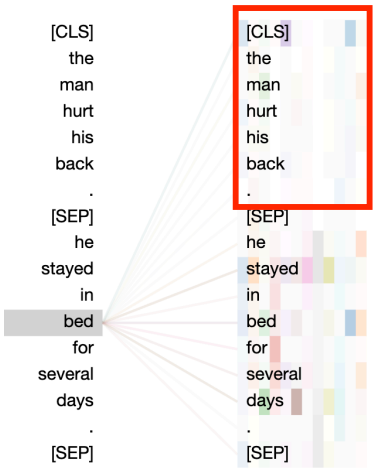

[CLS]
The
man
hurt
his
back
.
[SEP]
He
stayed
in
bed
for
several
days
.
[SEP]
[CLS]
The
man
hurt
his
back
.
[SEP]
He
stayed
in
bed
for
several
days
.
[SEP]
[CLS]
.
[SEP]
He
stayed
in
bed
for
several
days
.
[SEP]
[CLS]
.
[SEP]
He
stayed
in
bed
for
several
days
.
[SEP]

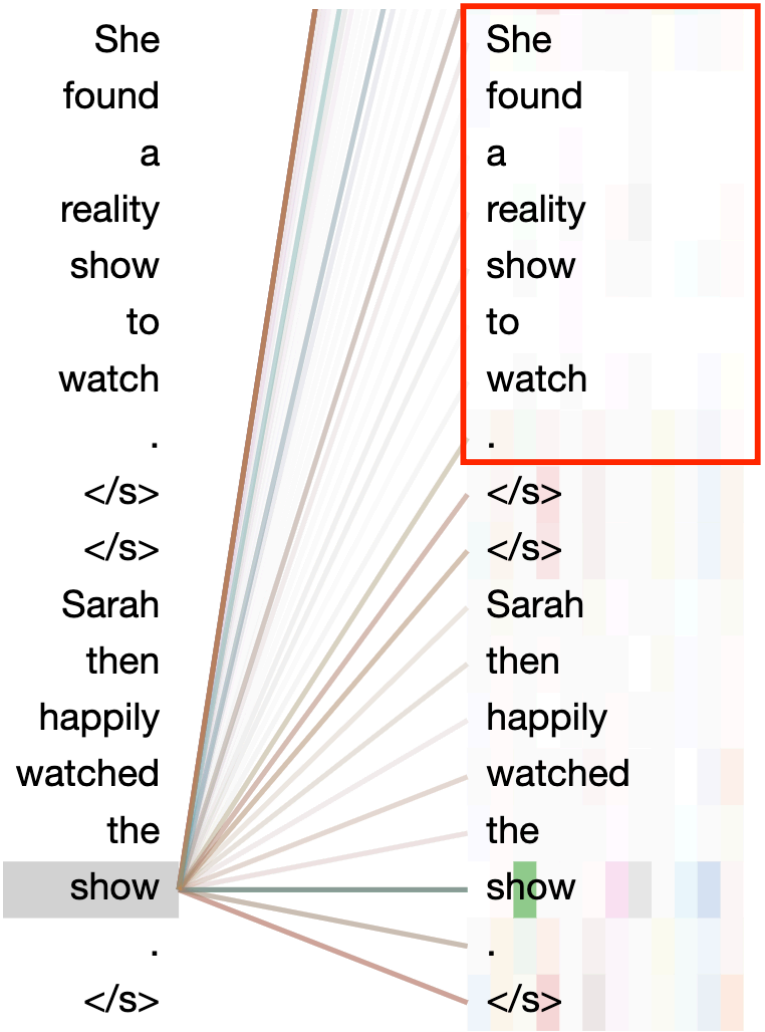

[CLS]
The
man
hurt
his
back
.
[SEP]
He
stayed
in
bed
for
several
days
.
[SEP]
[CLS]
The
man
hurt
his
back
.
[SEP]
He
stayed
in
bed
for
several
days
.
[SEP]

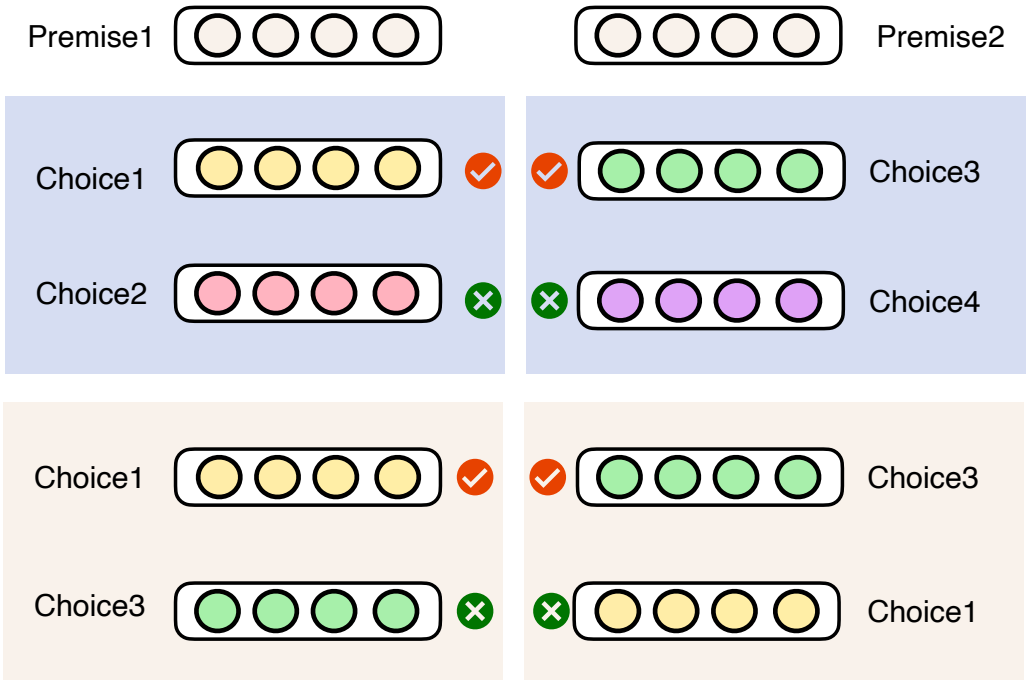

[CLS]
The
boy
failed
his
history
exam
.
[SEP]
He
forgot
to
study
.
[SEP]
[CLS]
The
boy
failed
his
history
exam
.
[SEP]
He
forgot
to
study
.
[SEP]
[CLS]
.
[SEP]
He
forgot
to
study
.
[SEP]
[CLS]
.
[SEP]
He
forgot
to
study
.
[SEP]

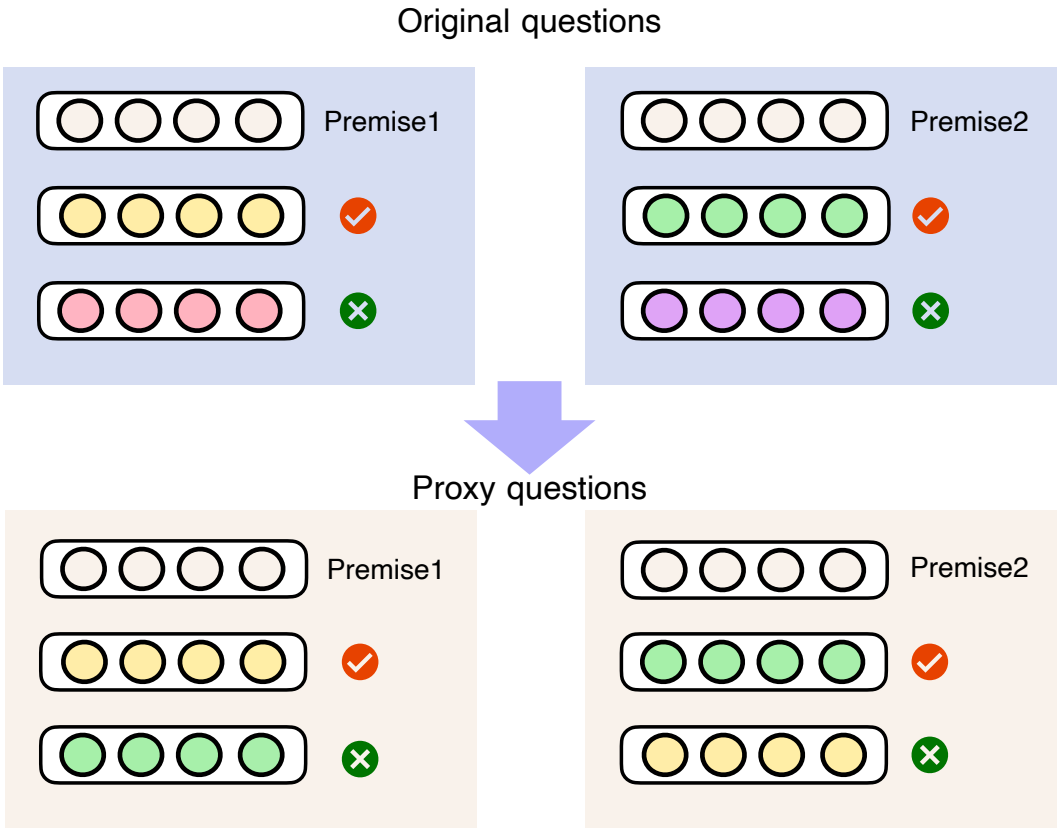

She
found
a
reality
show
to
watch
.
</s>
</s>
Sarah
then
happily
watched
the
show
.
</s>
She
found
a
reality
show
to
watch
.
</s>
</s>
Sarah
then
happily
watched
the
show
.
</s>

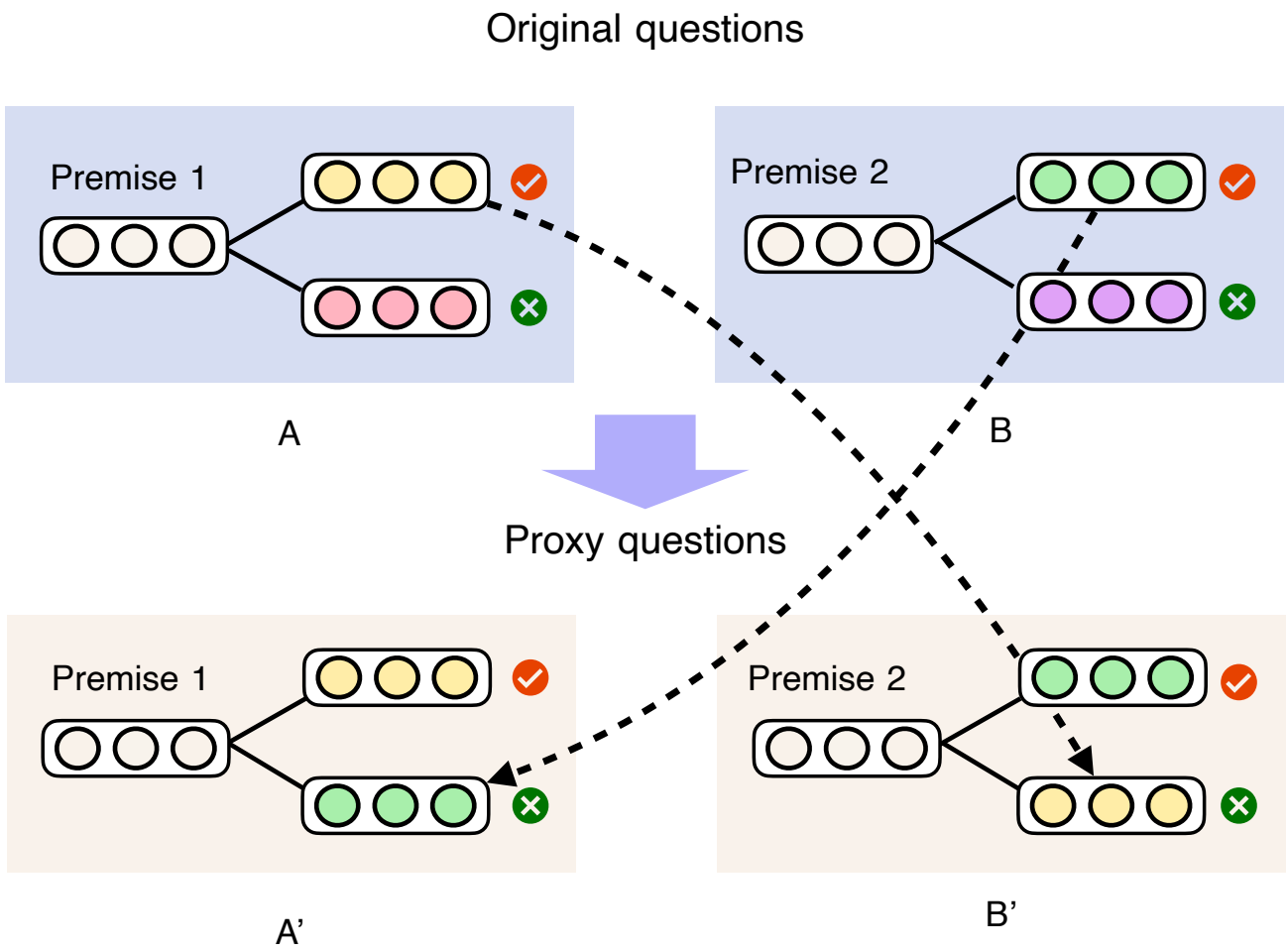

[CLS]
the
man
hurt
his
back
.
[SEP]
he
stayed
in
bed
for
several
days
.
[SEP]
[CLS]
the
man
hurt
his
back
.
[SEP]
he
stayed
in
bed
for
several
days
.
[SEP]

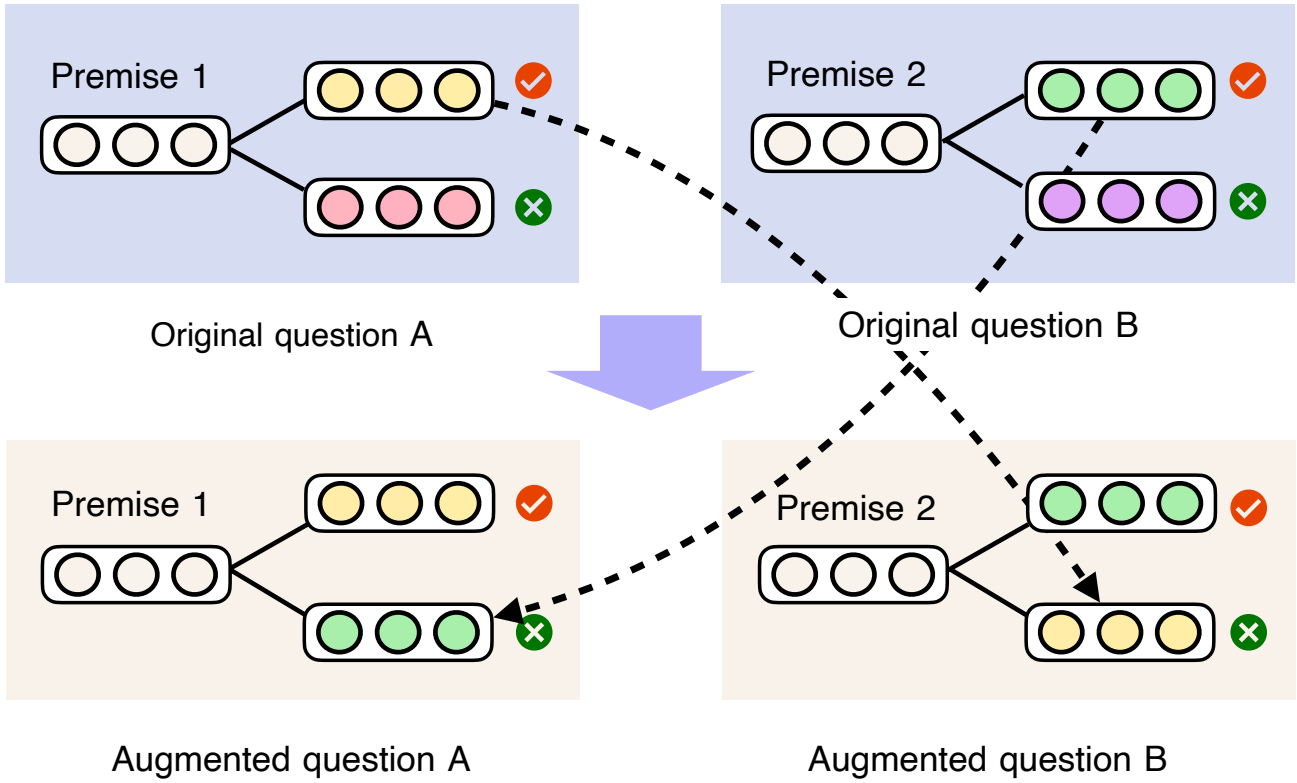

[CLS]
the
man
hurt
his
back
.
[SEP]
he
stayed
in
bed
for
several
days
.
[SEP]
[CLS]
the
man
hurt
his
back
.
[SEP]
he
stayed
in
bed
for
several
days
.
[SEP]

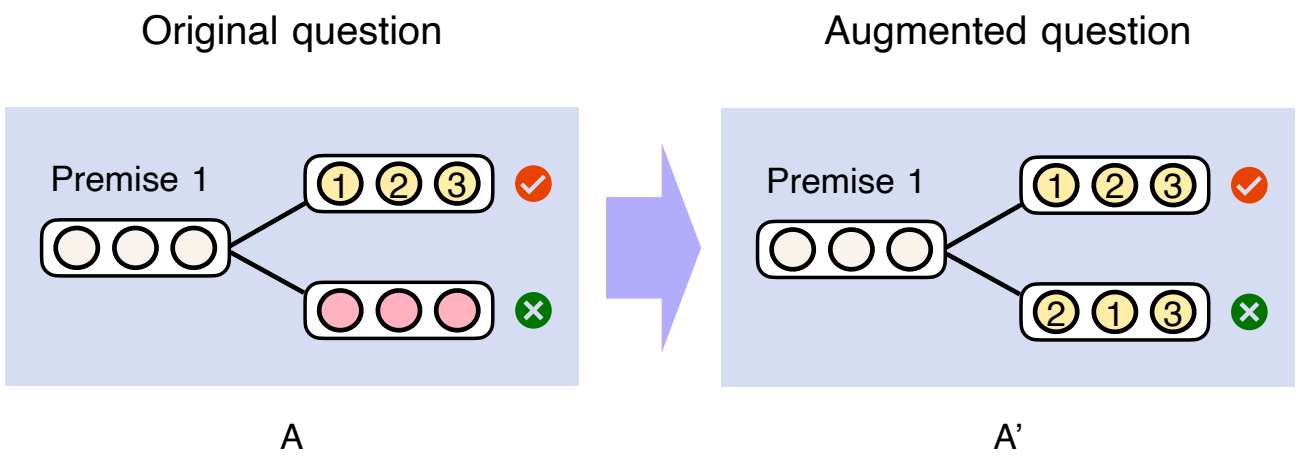

She
found
a
reality
show
to
watch
.
</s>
</s>
Sarah
then
happily
watched
the
show
.
</s>
She
found
a
reality
show
to
watch
.
</s>
</s>
Sarah
then
happily
watched
the
show
.
</s>

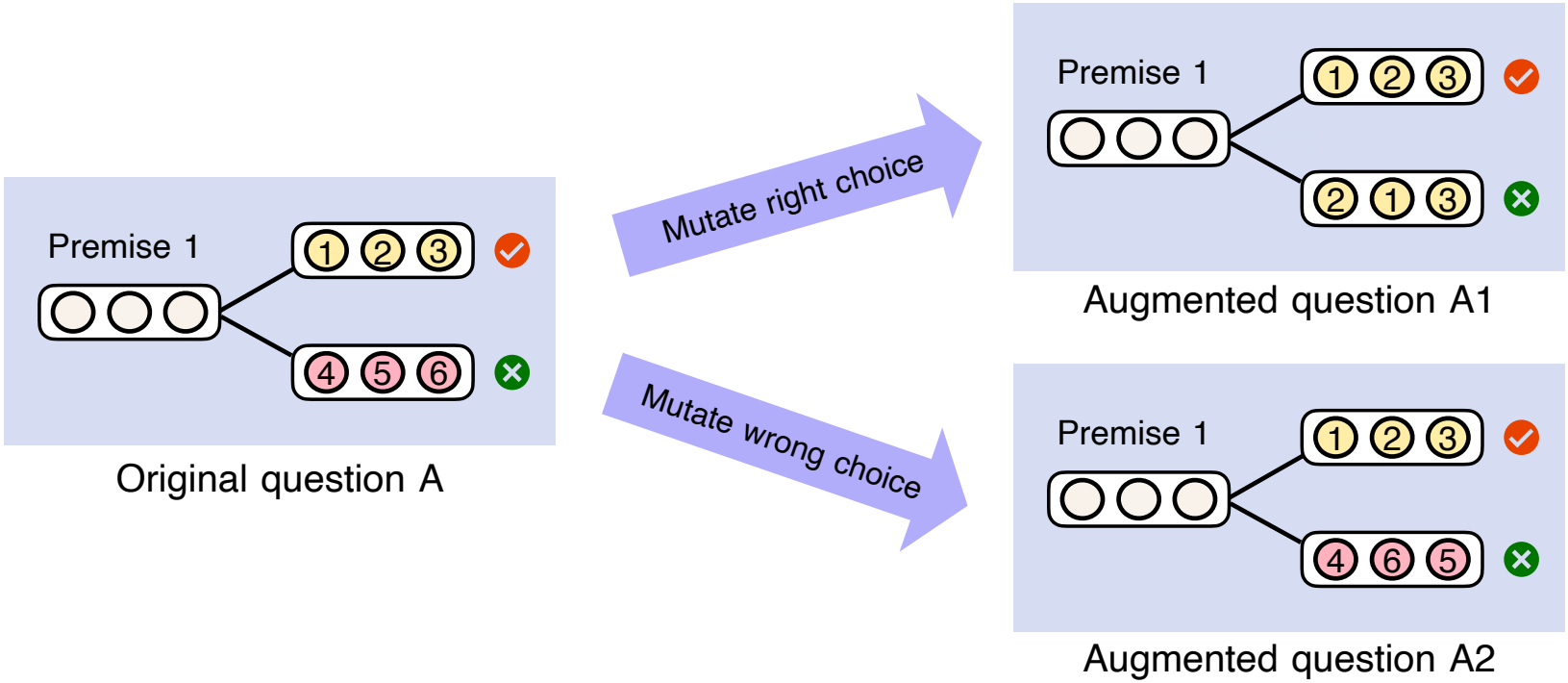

Premise1
Premise2
Choice1
Choice3
Choice2
Choice4
Choice1
Choice3
Choice3
Choice1

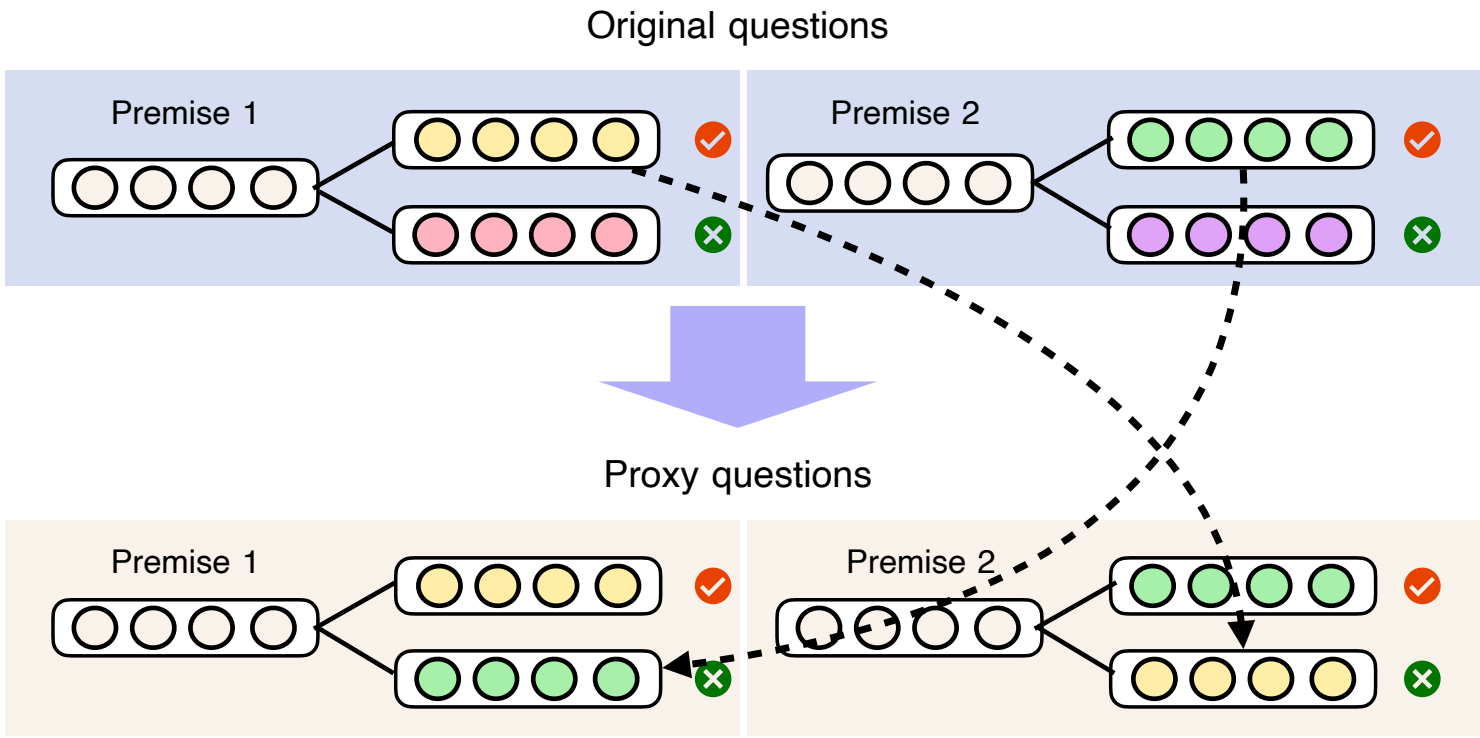

Original questions
Premise1
Premise2
Proxy questions
Premise1
Premise2

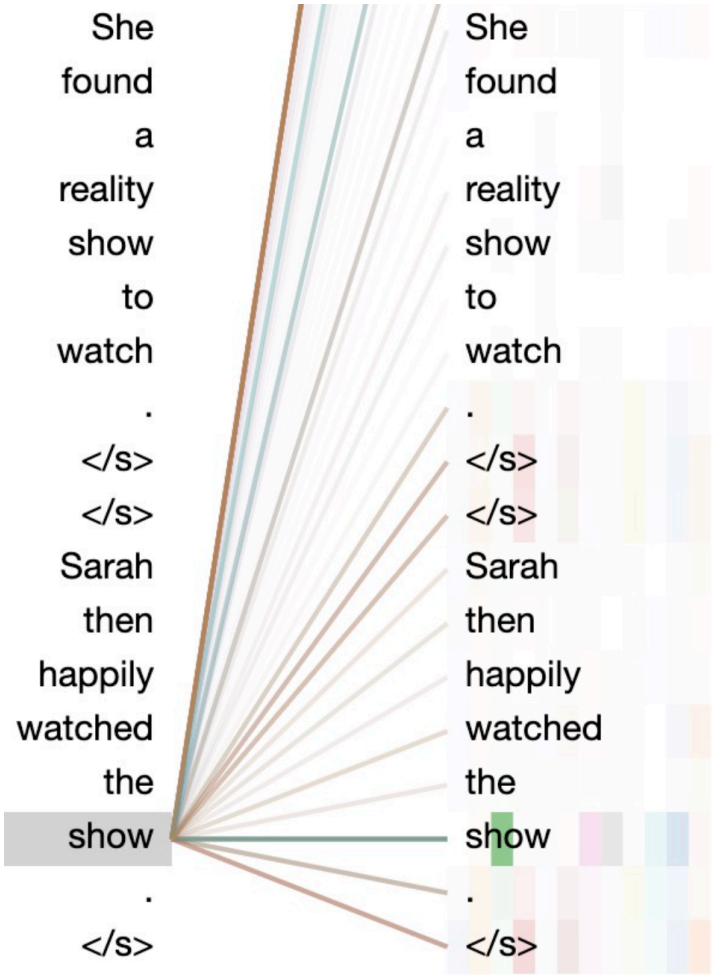

Original questions
Premise 1
Premise 2
A
B
Proxy questions
Premise 1
Premise 2
A’
B’

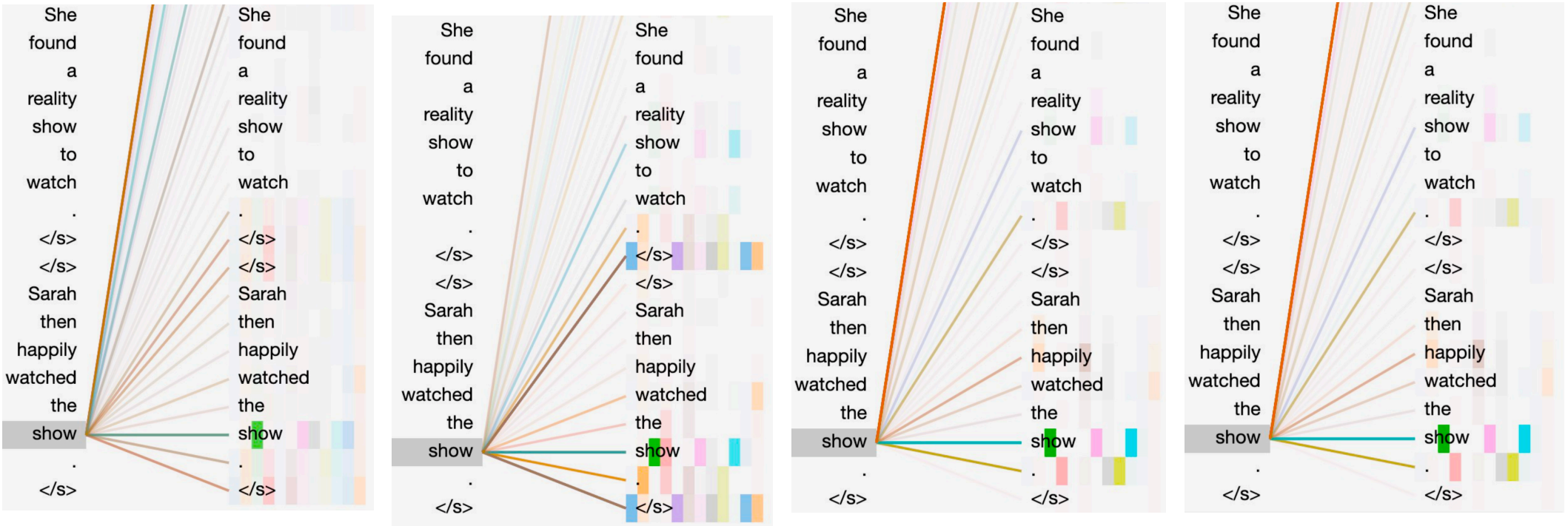

Premise 1
Premise 2
Original question A
Original question B
Premise 1
Premise 2
Augmented question A
Augmented question B

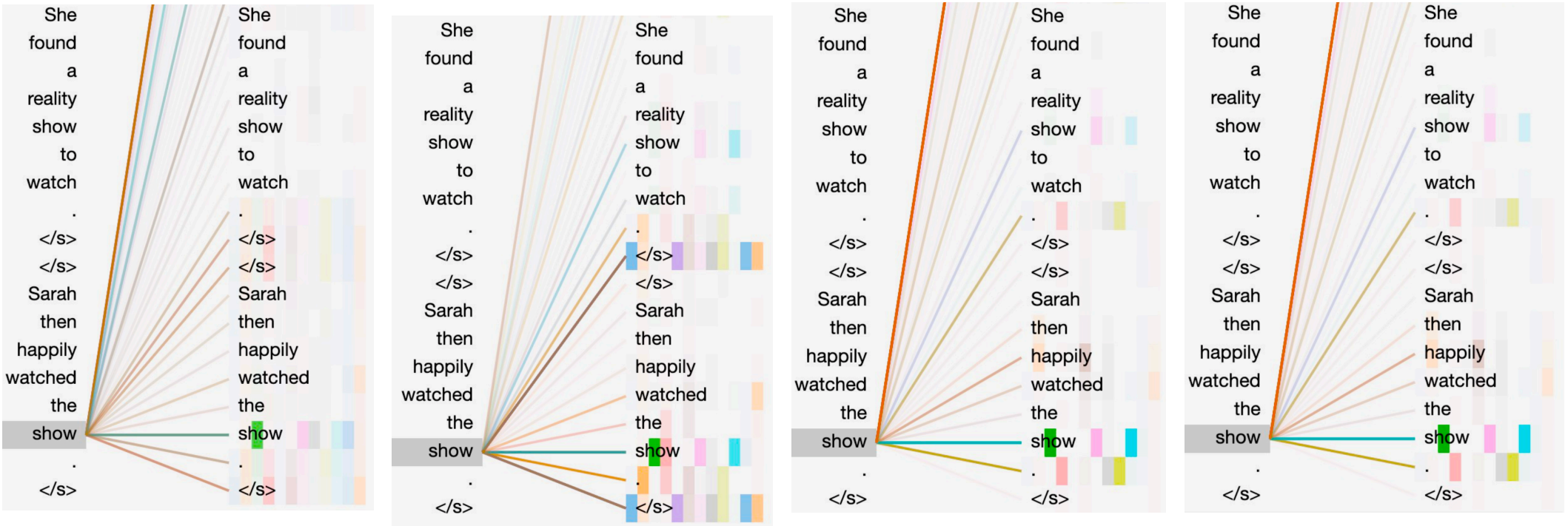

Original question
Augmented question
Premise 1
1 2 3
Premise 1
1 2 3
2 1 3
A
A’

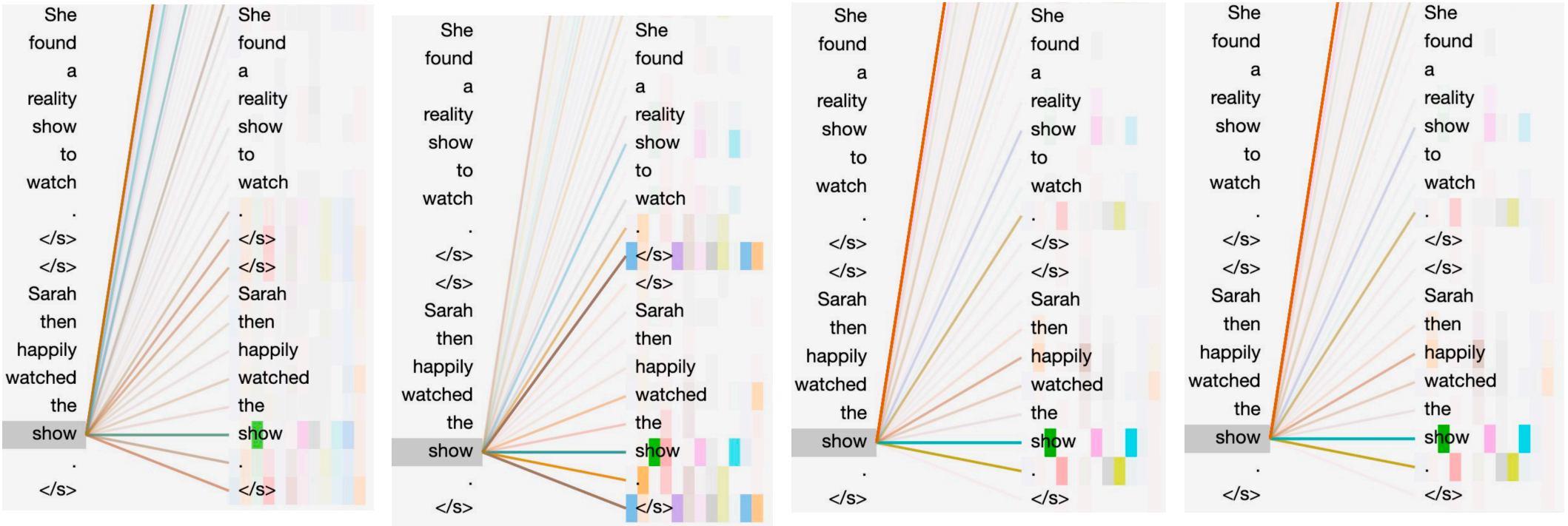

Premise 1
1 2 3
4 5 6
Original question A
Mutate right choice
Premise 1
1 2 3
2 1 3
Augmented question A1
Mutate wrong choice
Premise 1
1 2 3
4 6 5
Augmented question A2

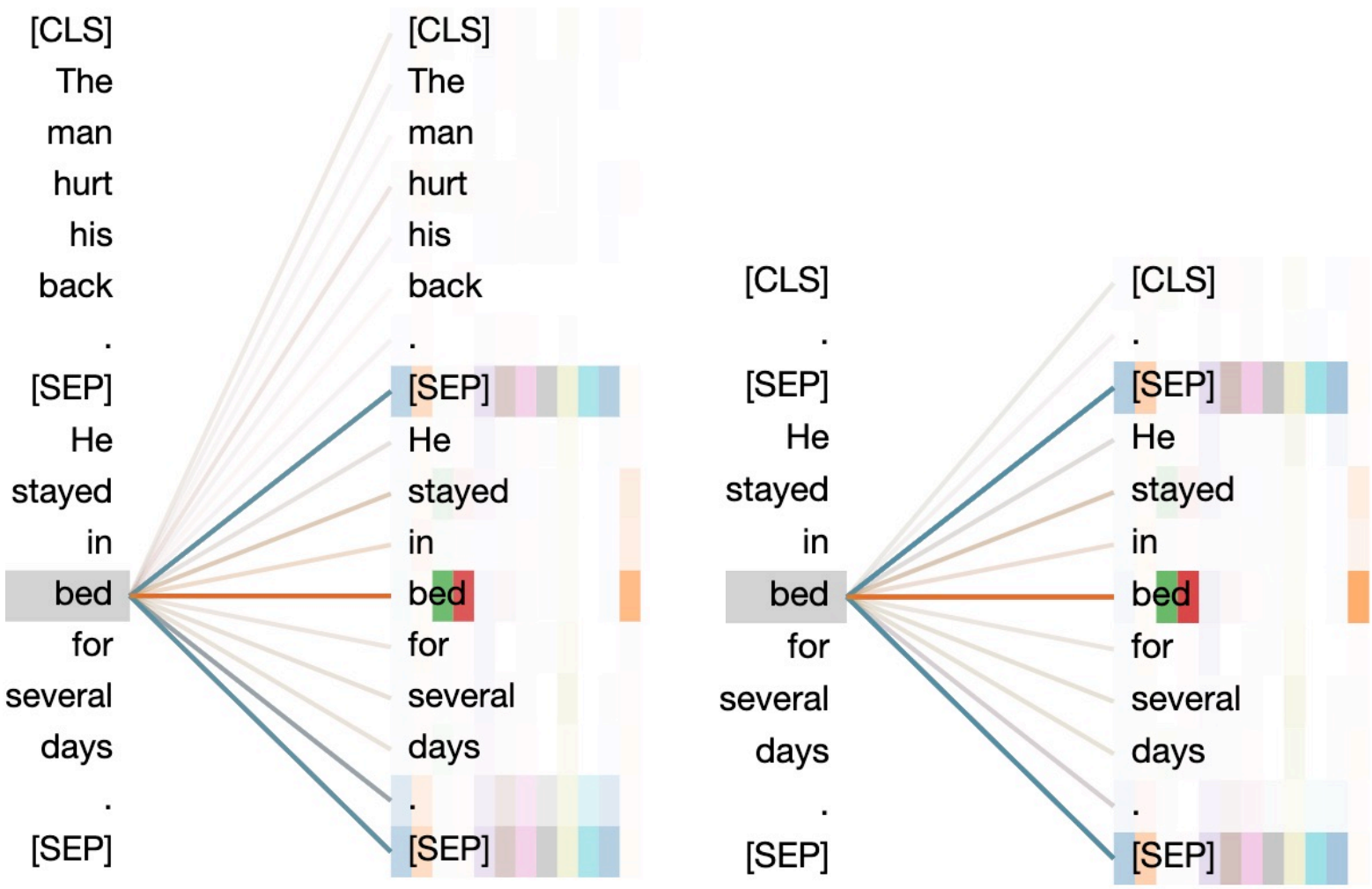

Original questions
Premise 1
Premise 2
Proxy questions
Premise 1
Premise 2

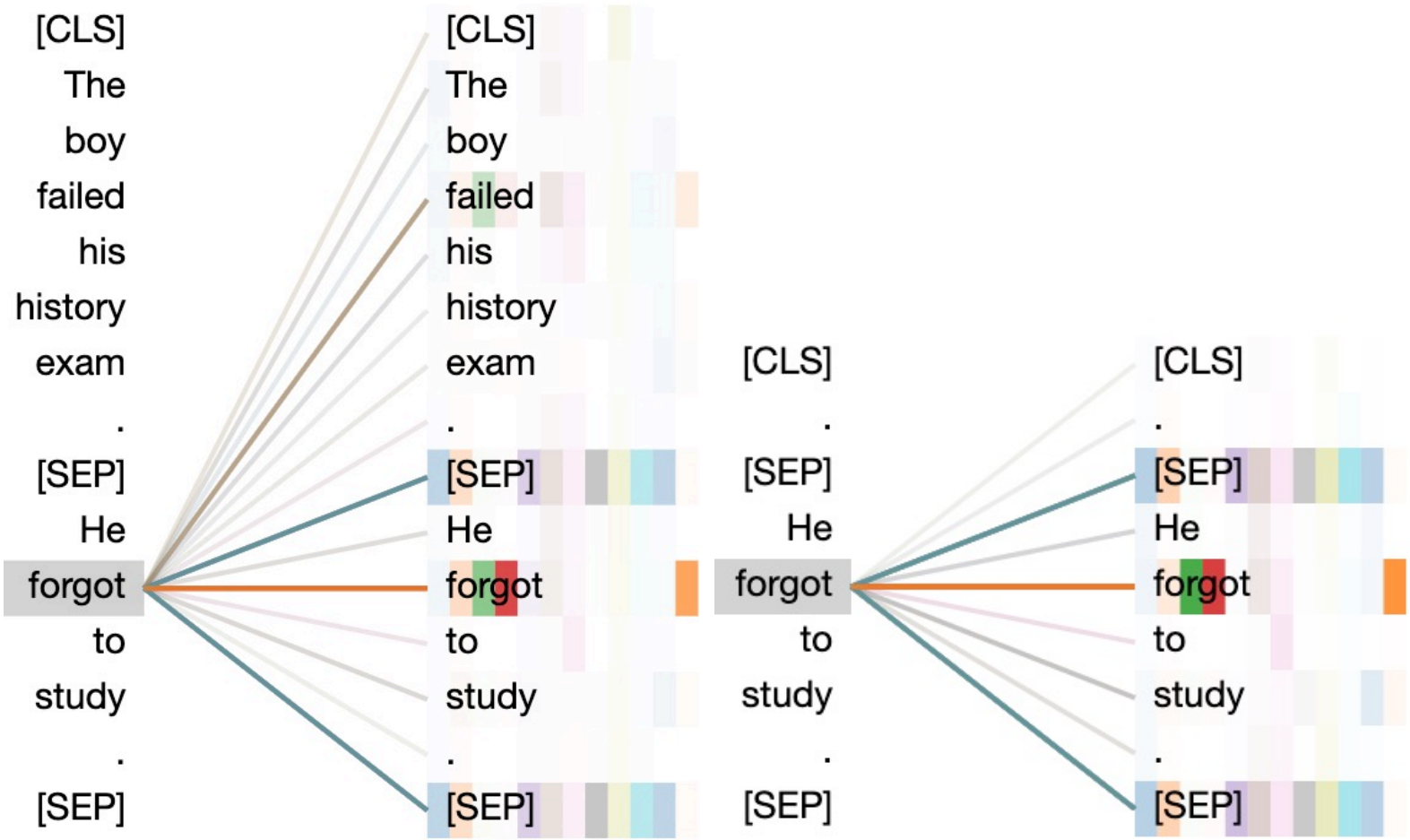

She
found
a
reality
show
to
watch
.
</s>
</s>
Sarah
then
happily
watched
the
show
.
</s>
She
found
a
reality
show
to
watch
.
</s>
</s>
Sarah
then
happily
watched
the
show
.
</s>

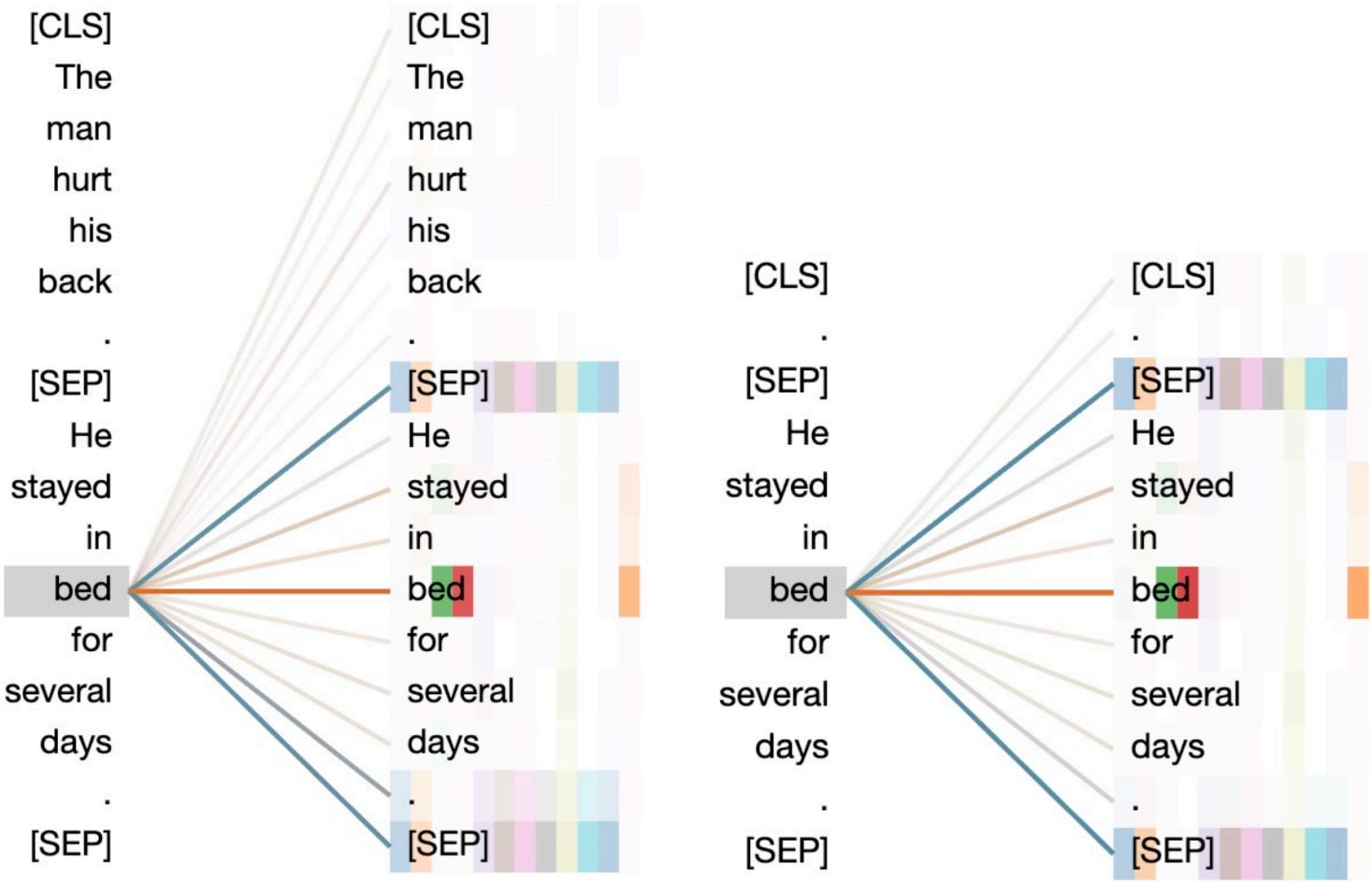

She found a reality show to watch . </s> </s> Sarah then happily watched the show . </s>
She found a reality show to watch . </s> </s> Sarah then happily watched the show . </s>

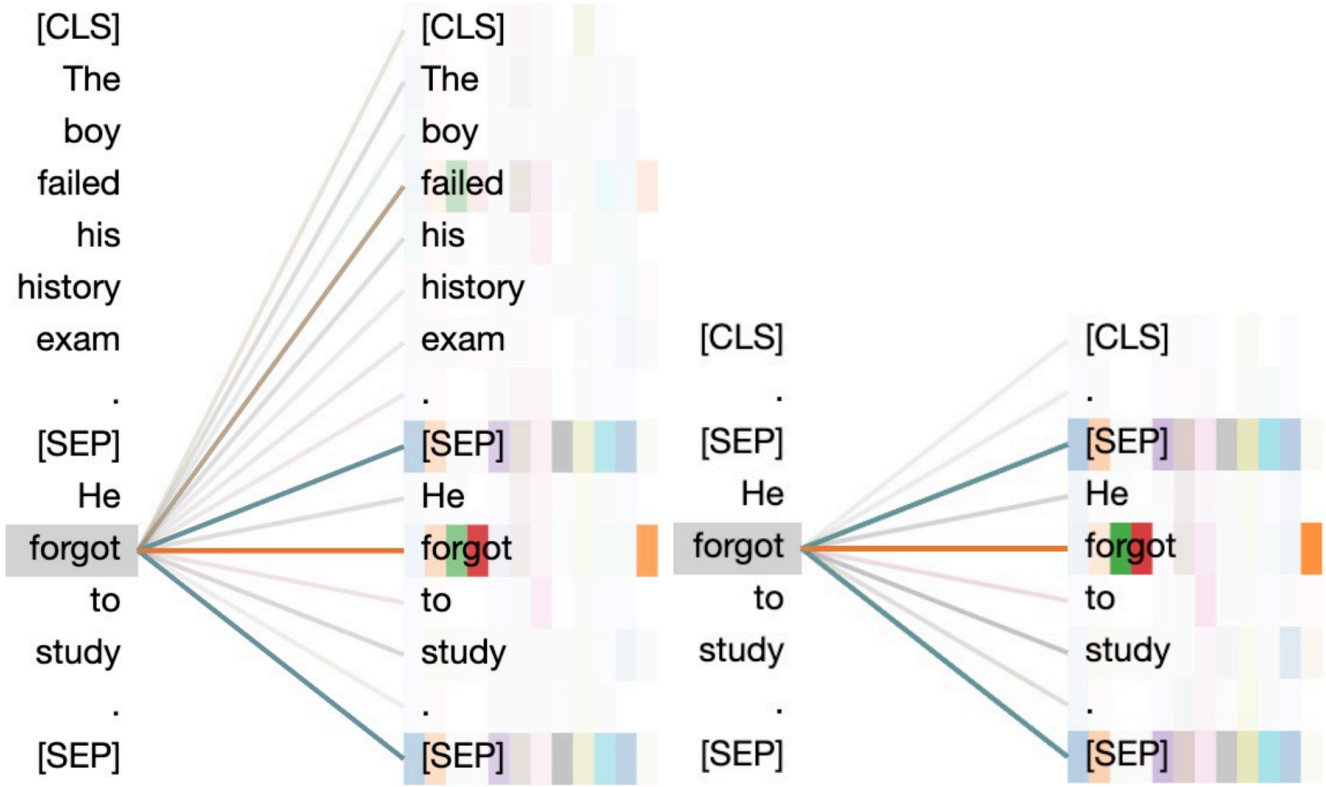

She found a reality show to watch . </s> </s> Sarah then happily watched the show . </s>
She found a reality show to watch . </s> </s> Sarah then happily watched the show . </s>


She found a reality show to watch . </s> </s> Sarah then happily watched the show . </s>
She found a reality show to watch . </s> </s> Sarah then happily watched the show . </s>


[CLS]
The
man
hurt
his
back
.
[SEP]
He
stayed
in
bed
for
several
days
.
[SEP]
[CLS]
The
man
hurt
his
back
.
[SEP]
He
stayed
in
bed
for
several
days
.
[SEP]
[CLS]
.
[SEP]
He
stayed
in
bed
for
several
days
.
[SEP]
[CLS]
.
[SEP]
He
stayed
in
bed
for
several
days
.
[SEP]


[CLS]
The
boy
failed
his
history
exam
.
[SEP]
He
forgot
to
study
.
[SEP]
[CLS]
The
boy
failed
his
history
exam
.
[SEP]
He
forgot
to
study
.
[SEP]
[CLS]
.
[SEP]
He
forgot
to
study
.
[SEP]
[CLS]
.
[SEP]
He
forgot
to
study
.
[SEP]


[CLS]
The
man
hurt
his
back
.
[SEP]
He
stayed
in
bed
for
several
days
.
[SEP]
[CLS]
The
man
hurt
his
back
.
[SEP]
He
stayed
in
bed
for
several
days
.
[SEP]
[CLS]
.
[SEP]
He
stayed
in
bed
for
several
days
.
[SEP]
[CLS]
.
[SEP]
He
stayed
in
bed
for
several
days
.
[SEP]


[CLS]
The
boy
failed
his
history
exam
.
[SEP]
He
forgot
to
study
.
[SEP]
[CLS]
The
boy
failed
his
history
exam
.
[SEP]
He
forgot
to
study
.
[SEP]
[CLS]
.
[SEP]
He
forgot
to
study
.
[SEP]
[CLS]
.
[SEP]
He
forgot
to
study
.
[SEP]

Probing spurious features

Evaluate Dataset

Evaluate Model

For this dataset:
The vocabulary size in training data is: ?
The vocabulary size in test data is : ?
Accuracy for our linear model is :?

Spurious features:

- [ ] Word
- [ ] Overlap
- [ ] Typos
- [ ] Sentiment
- [ ] Negation
- [ ] POS
- [ ] NER

Submit

Probing spurious features

Evaluate Dataset

Evaluate Model

Spurious features: Words

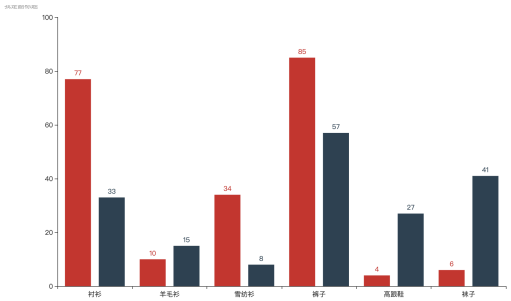
Second level feature: top 10 words:?

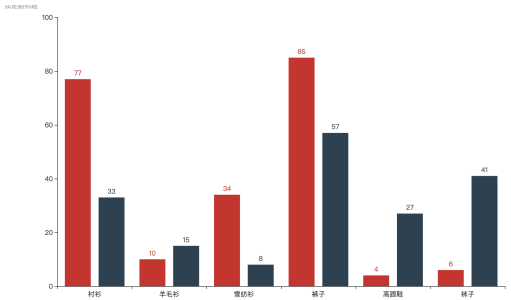
Train

Test

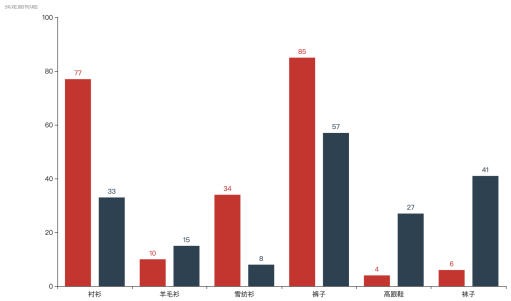
KL score

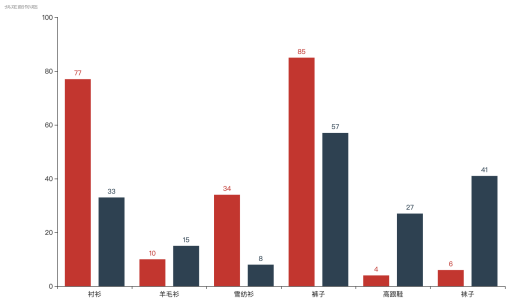
Word : nobody





Word : home





Probing spurious features

Evaluate Dataset

Evaluate Model

Please upload your dataset with the following format :

1. The data file should be in txt data format.
2. The data should contain four parts: ID, PREDICTION on each line and connected with “\t”

For example: 0\tentailement\n

Test prediction file Upload 在这里加一个判定，如果没有upload dataset 显示没有数据集，然后跳回1

Submit

Probing spurious features

Evaluate Dataset

Evaluate Model

For this model:
Accuracy is: ?
Our baseline accuracy is: ?

Spurious features:

- [ ] Word
- [ ] Overlap
- [ ] Typos
- [ ] Sentiment
- [ ] Negation
- [ ] POS
- [ ] NER

Submit

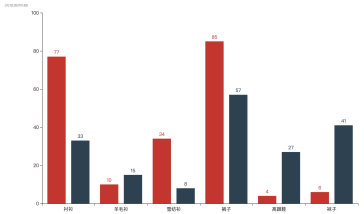
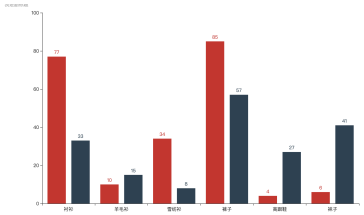
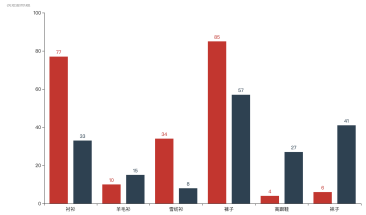
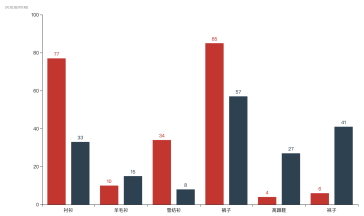
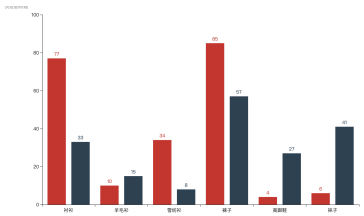
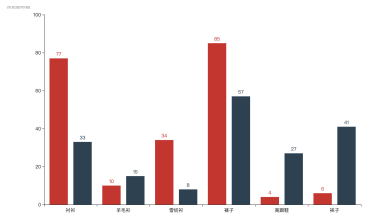
Probing spurious features

Evaluate Dataset

Evaluate Model

Spurious features: Words

Second level feature: top 10 words:?

	Train	Test	Prediction	KL score
Word : nobody				
Word : home				

Probing spurious features

Evaluate Dataset

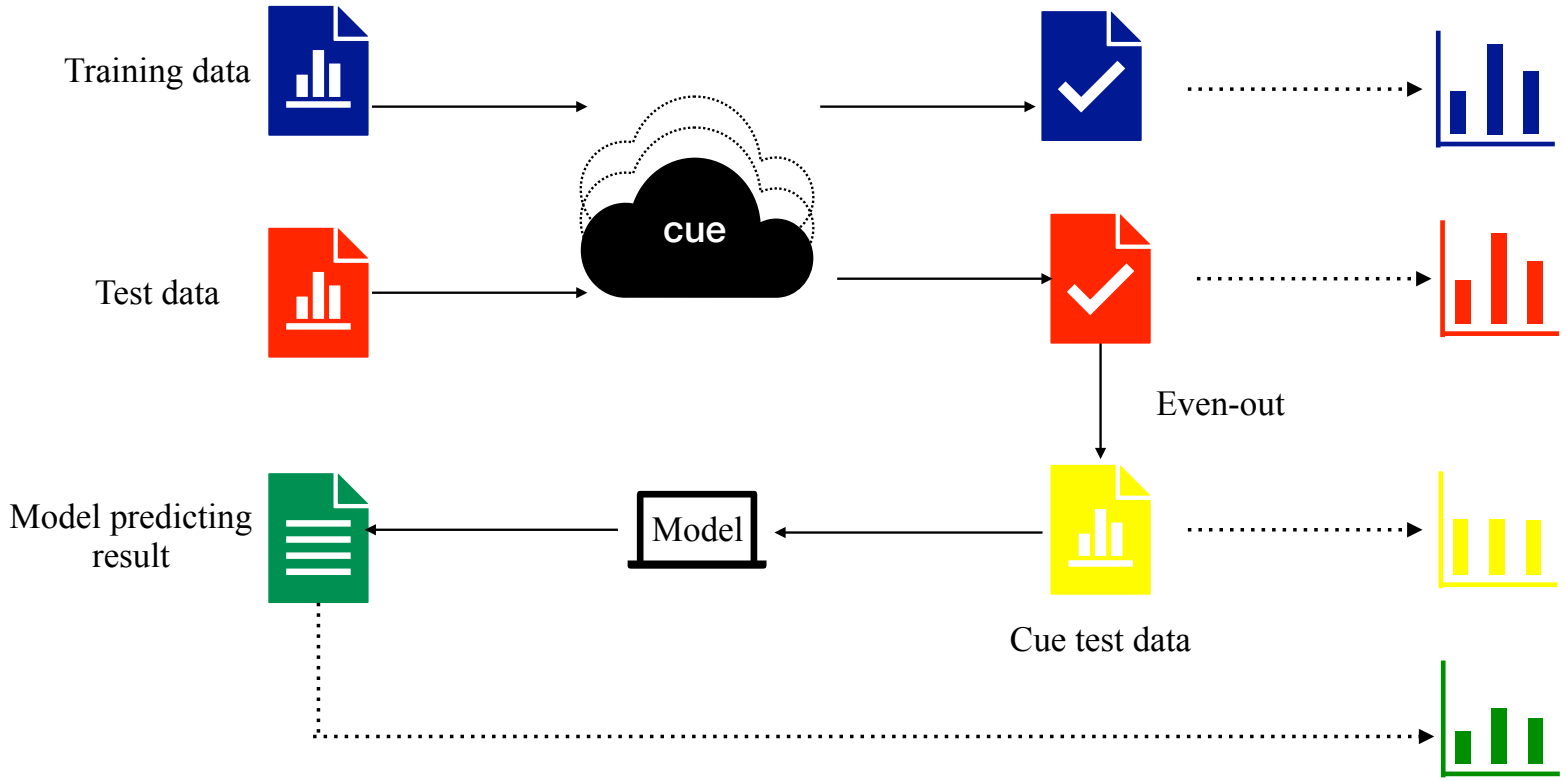
Evaluate Model

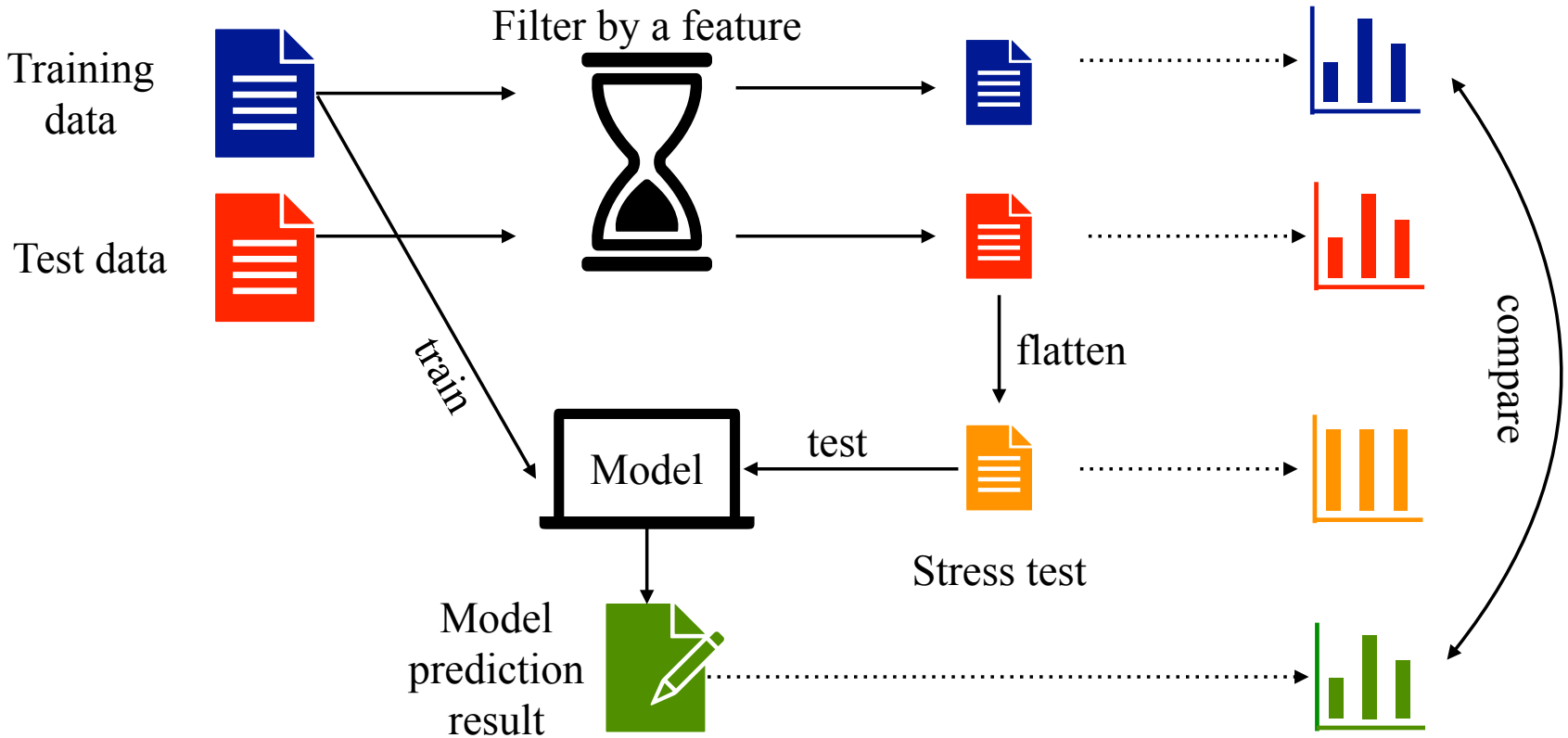
For this model:
Accuracy is: ?
Our baseline accuracy is: ?

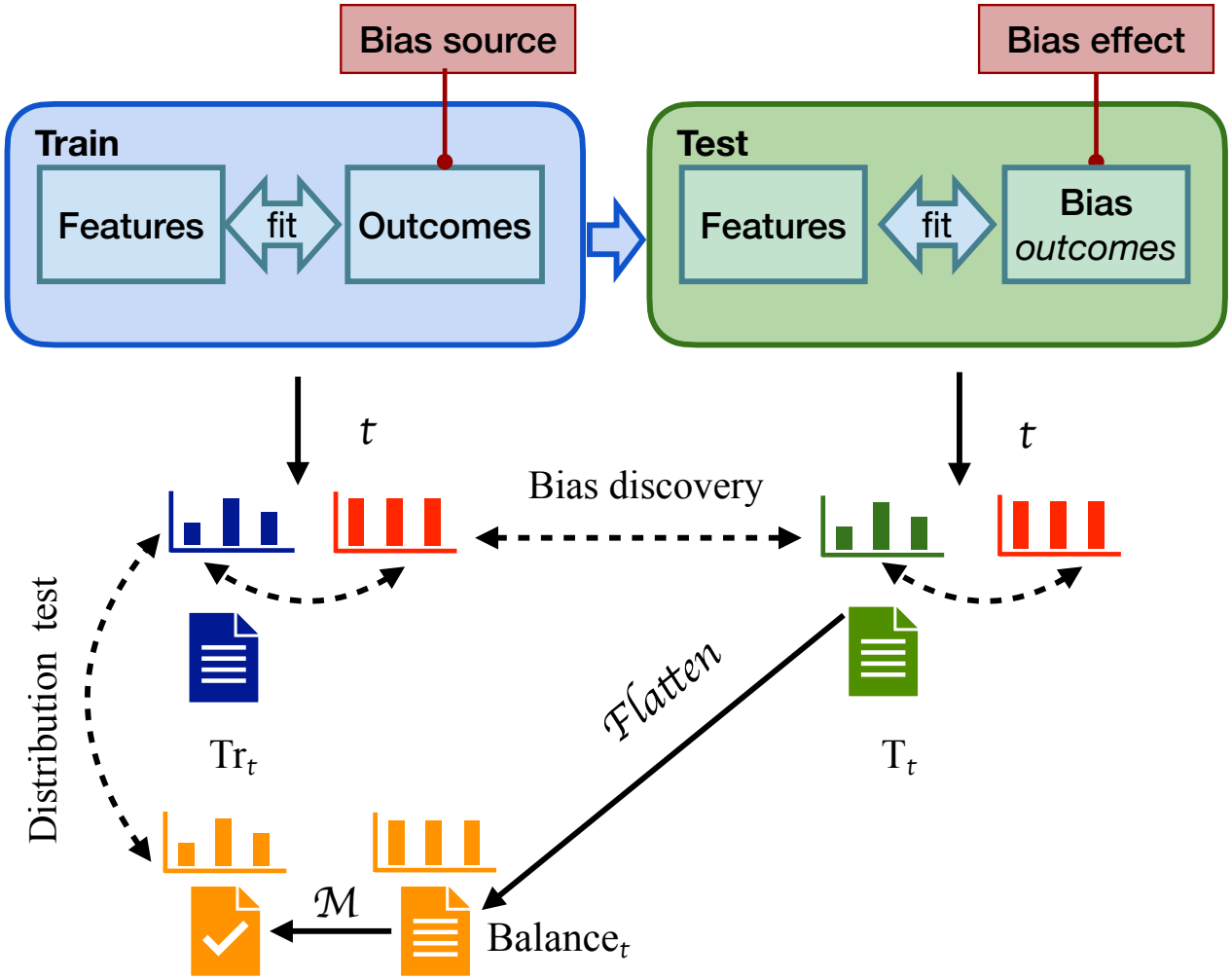
Spurious features:

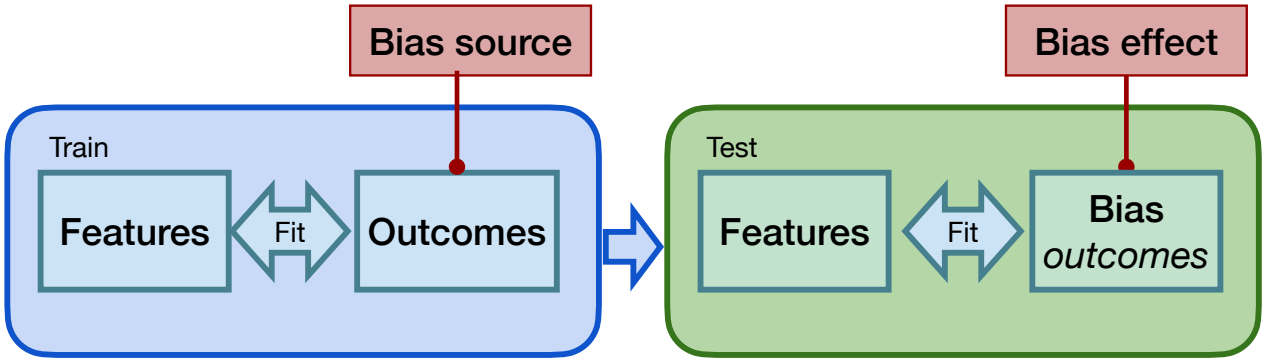
- [ ] Word
- [ ] Overlap
- [ ] Typos
- [ ] Sentiment
- [ ] Negation
- [ ] POS
- [ ] NER

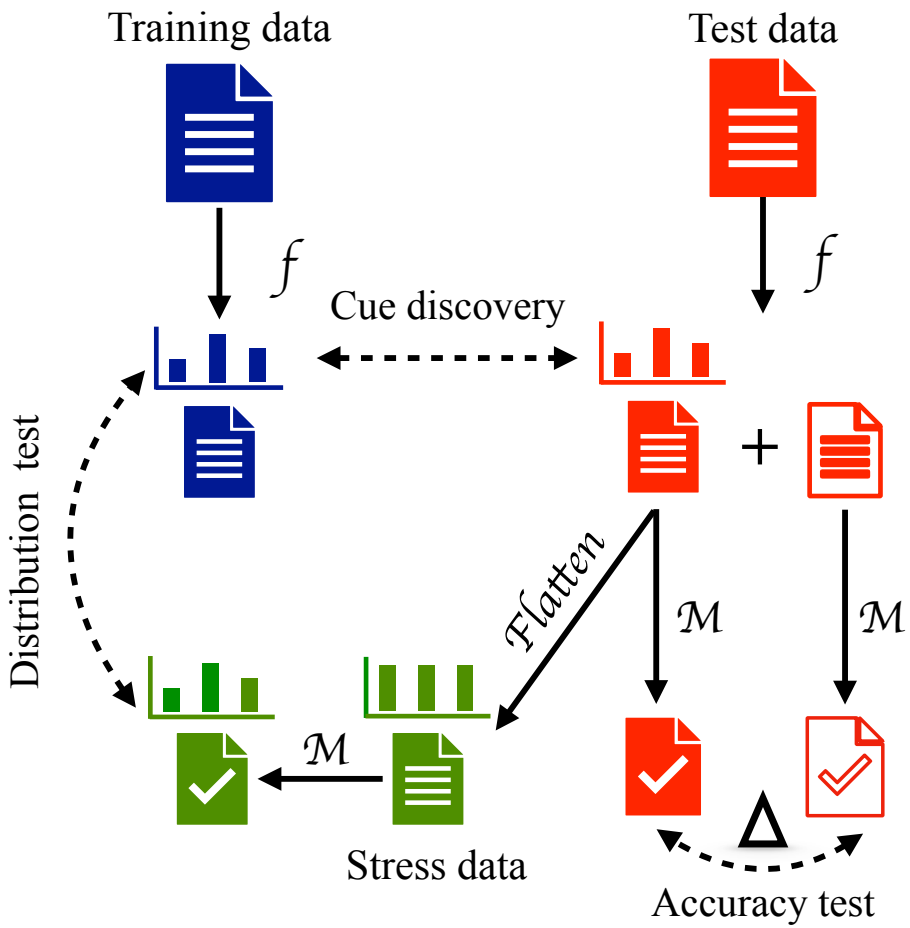
Submit

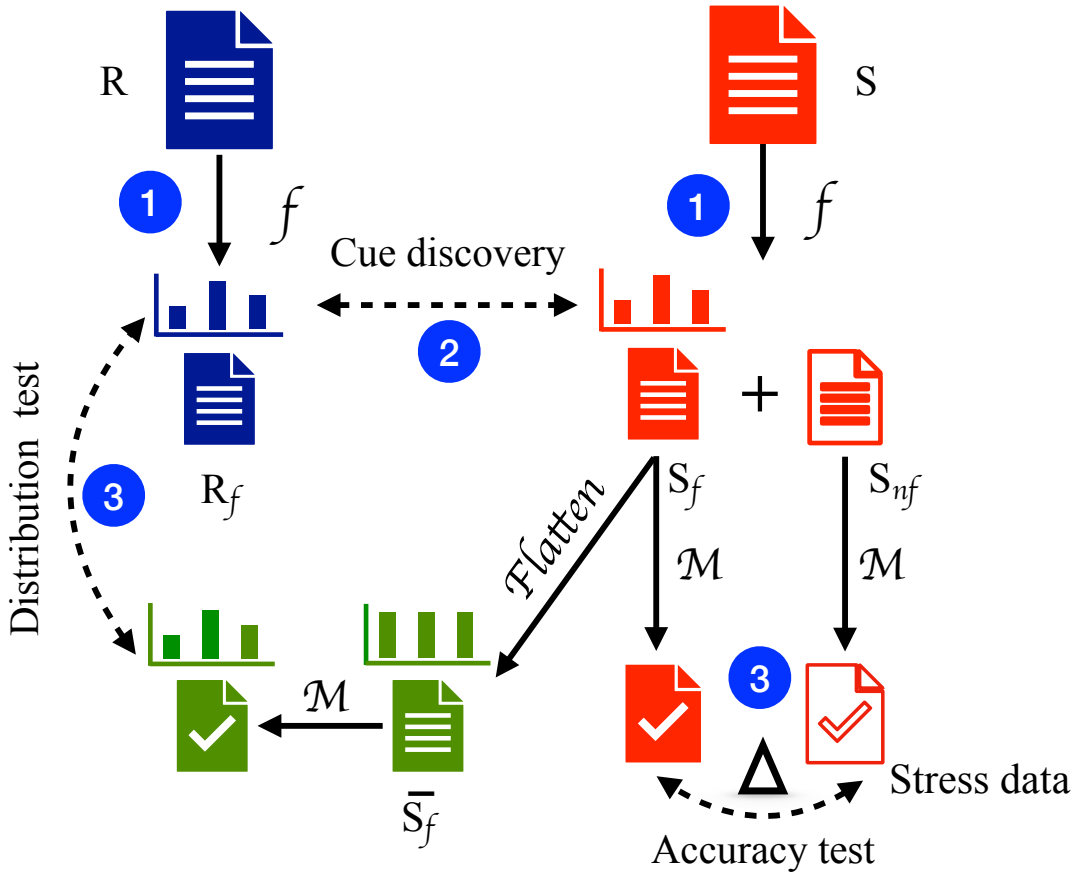

Training data
Test data
cue
Even-out
Model predicting result
Model
Cue test data

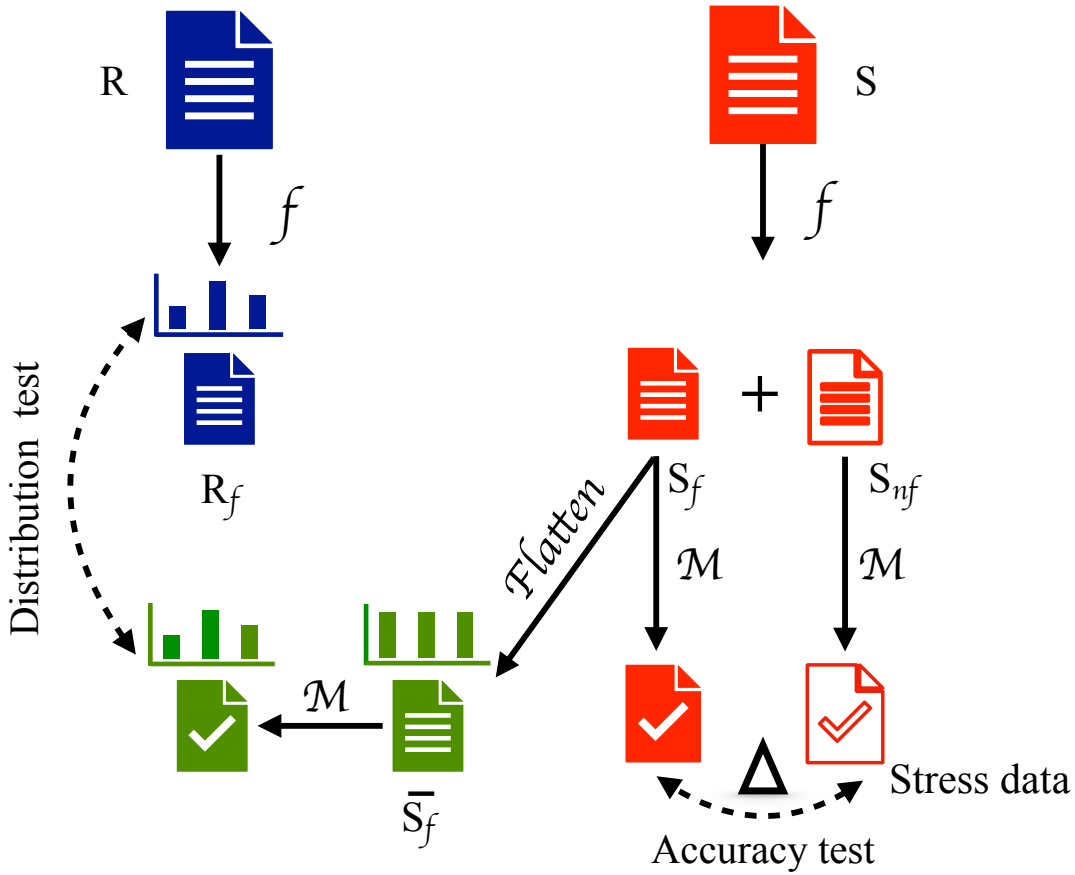

Training data
Test data
Filter by a feature
train
flatten
Model
test
Stress test
Model prediction result
compare


Bias source
Bias effect
Train
Features
fit
Outcomes
Test
Features
fit
Bias
outcomes
t
t
Bias discovery
Distribution test
Tr$_t$
T$_t$
Flatten
$\mathcal{M}$
Balance$_t$


Bias source
Bias effect
Train
Features
Fit
Outcomes
Test
Features
Fit
Bias
outcomes


Training data
Test data
f
f
Cue discovery
+
Distribution test
Flatten
M
M
M
Stress data
Δ
Accuracy test


R
S
1
f
1
f
Cue discovery
2
Distribution test
3
R_f
S_f
+
S_{nf}
Flatten
$\mathcal{M}$
$\mathcal{M}$
$\mathcal{M}$
3
$\bar{S}_f$
Δ
Stress data
Accuracy test


R
S
f
f
Distribution test
R_f
S_f
S_{nf}
Flatten
$\mathcal{M}$
$\mathcal{M}$
$\mathcal{M}$
$\bar{S}_f$
Δ
Stress data
Accuracy test

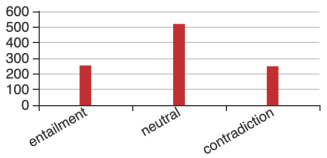
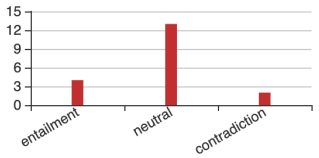
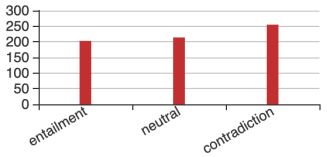
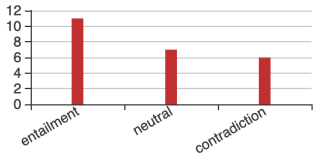
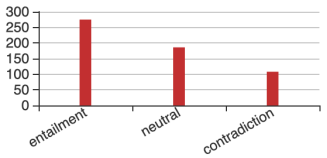
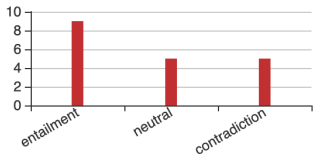
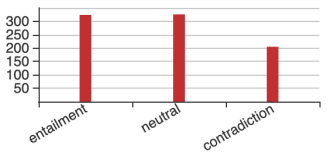
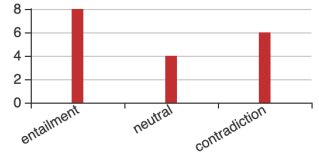
Probing spurious features

Evalute Dataset

Evalute Model

Spurious features: word

Second level feature: filled uniforms books giving pushing speaking sunglasses kneeling signs surrounded

	Train	Test	KL score
giving			0.09393793315400147
pushing			0.06039326965589334
speaking			0.018589477685614704
sunglasses			0.06643377287113857

Probing spurious features

If you don't have your own dataset, you can choose a popular dataset in the following:

○ SNLI ○ QNLI ○ MNLI ○ ROC ○ COPA ○ SWAG ○ ARCT ○ ARCT_adv
○ RACE ○ RECLOR ○ CQA ○ Ubuntu

If you want to test yout own dataset, please upload your dataset with the following format:

1. The data file should be in csv data format.

2. The data should contain four parts: ID, PREMISE, HYPOTHESIS and LABEL.

For example:

ID	PREMISE	HYPOTHESIS	LABEL
0	A soccer game with multiple males playing.	Some men are playing a sport.	Entailment

Training data file Choose File no file selected

Test data file Choose File no file selected

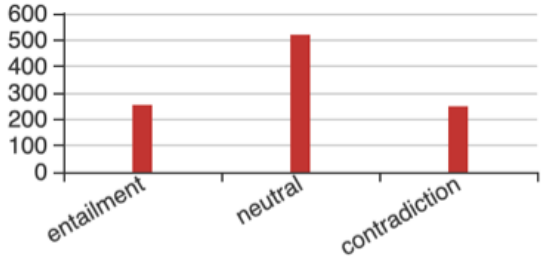
Evalute Dataset

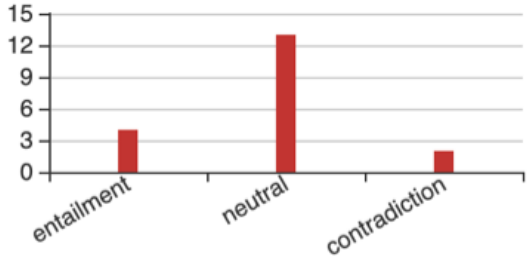
Please upload your model prediction result on test data with the following format:

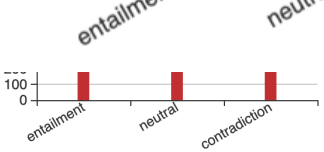
1. The data file should be in txt data format.

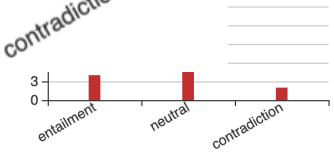
2. The data should contain four parts: ID, PREDICTION on each line and connected with "\t".

For example: 0\tentailement\n



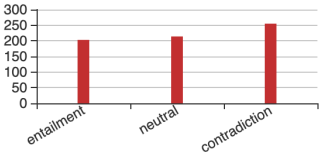


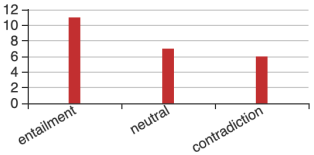




0.09393793315400147

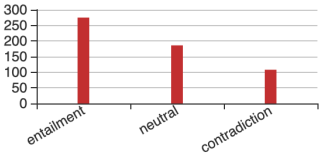
pushing

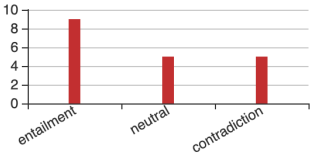




0.06039326965589334

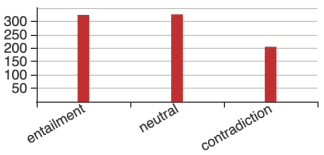
speaking

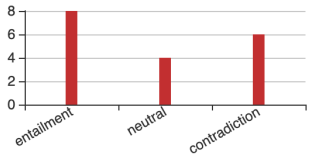




0.018589477685614704

sunglasses





0.06643377287113857

Probing spurious features

Spurious features:

- [ ] Word
- [ ] Overlap
- [ ] Typos
- [ ] Sentiment
- [ ] Negation
- [ ] POS
- [ ] NER

If you don't have your own dataset, you can choose a popular dataset in the following:

( ) SNLI ( ) QNLI ( ) MNLI ( ) ROC ( ) COPA ( ) SWAG ( ) ARCT ( ) ARCT_adv
( ) RACE ( ) RECLOR ( ) CQA ( ) Ubuntu

If you want to test yout own dataset, please upload your dataset with the following format:

1. The data file should be in csv data format.

2. The data should contain four parts: ID, PREMISE, HYPOTHESIS and LABEL.

For example:

ID	PREMISE	HYPOTHESIS	LABEL
0	A soccer game with multiple males playing.	Some men are playing a sport.	Entailment

Training data file Choose File | no file selected

Test d **Please choose a model to test in the following:**

( ) Bert ( ) ESIM ( ) fastText

Evalute Dataset

Please choose a model to test in the following:

( ) Bert ( ) ESIM ( ) fastText

Please upload your model prediction result on test data with the following format:

1. The data file should be in txt data format.

2. The data should contain four parts: ID, PREDICTION on each line and connected with "\t".

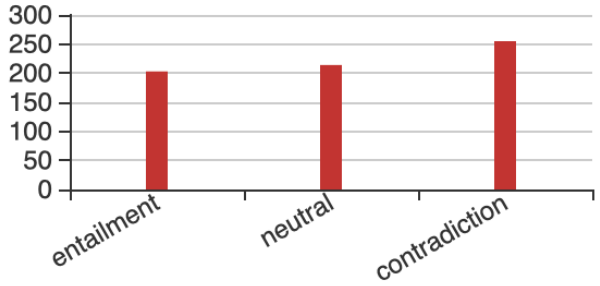
For example: 0\tentailement\n

Test prediction file Choose File | no file selected

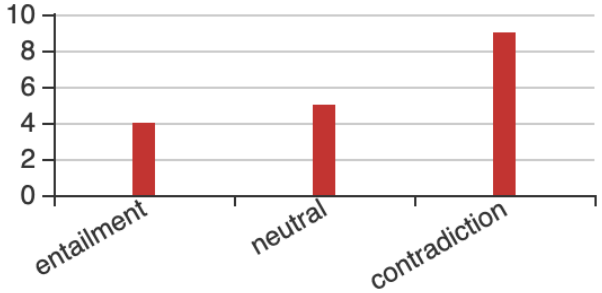
If you have test prediction file.

Evalute Model

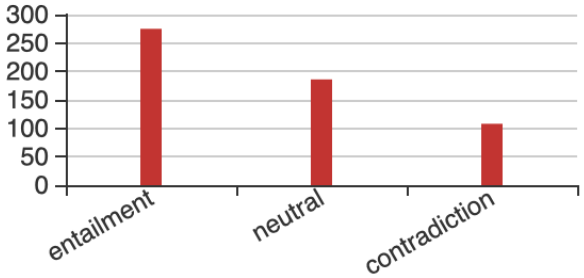
Train

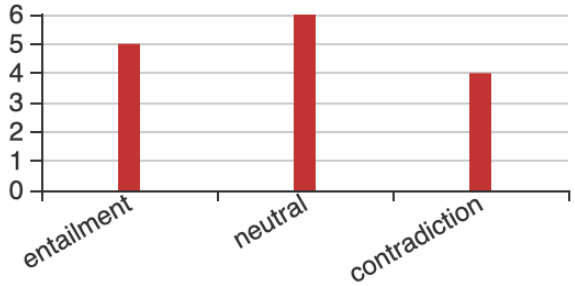


Prediction

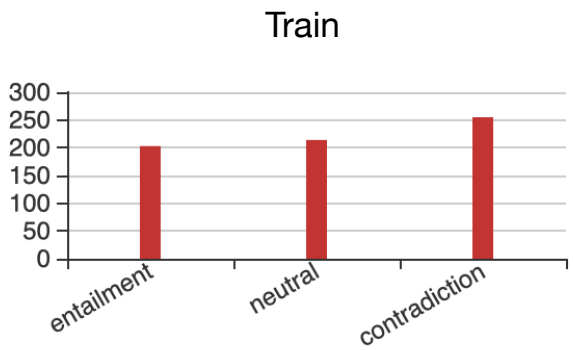


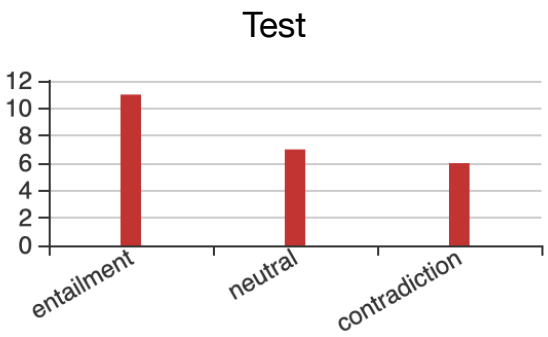
Feature = “pushing” JSD(Train, Prediction) = 0.0079



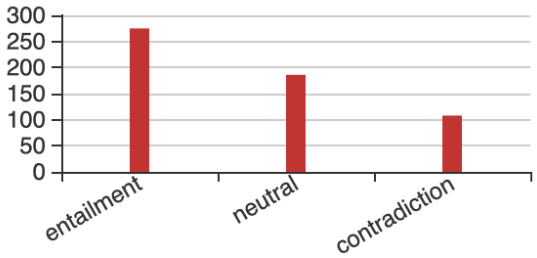


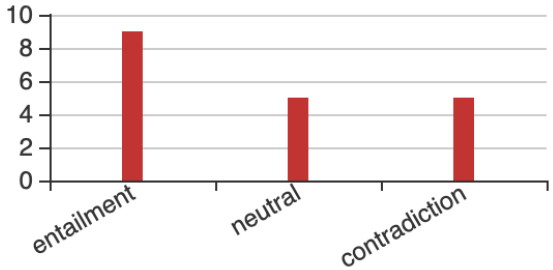
Feature = “speaking” JSD(Train, Prediction) = 0.0120





Feature = “pushing” MSE(Train) = 6.7053 JSD(Train, Test) = 0.0151





Feature = “speaking” MSE(Train) =73.6274 JSD(Train, Test) = 0.0017

Neg: train[1213,1384,3380] test:[20,28,49]pre:[12,24,24]
kl:0.0096 0.0722
Mean 1457.1099

overlap: trai[178605,175255,170281] test:[3304,3087,3014]pre:[3088,3002,2952]

0.0002 1.314e-05
200.80

```
'entailment': 203, 'neutral': 214, 'contradiction': [redacted]255

'entailment': 11, 'neutral': 7, 'contradiction': 6    }   4,5,9

马上到! :6.7053
Jsd: 0.0151
Jsd:0.0079

'entailment': 275, 'neutral': 186, 'contradiction': [redacted]108

'entailment': 9, 'neutral': 5, 'contradiction': 5     5,6,4
马上到! :73.6274
Jsd: 0.0017
Jsd:0.0120
```

ICQ (“I-see-cue”)

Visualizing statistical biases in Text Inference Datasets and Models

Spurious features: ☐ Word ☐ Overlap ☐ Typos ☐ Sentiment ☐ Negation ☐ POS ☐ NER

If you don’t have your own dataset, you can choose a popular dataset in the following:

○ SNLI ○ QNLI ○ MNLI ○ ROC ○ COPA ○ SWAG ○ ARCT ○ ARCT_adv
○ RACE ○ RECLOR ○ CQA ○ Ubuntu

If you want to test yout own dataset, please upload your dataset with the following format:

1. The data file should be in csv data format.

2. The data should contain four parts: ID, PREMISE, HYPOTHESIS and LABEL.

For example:

ID	PREMISE	HYPOTHESIS	LABEL
0	A soccer game with multiple males playing.	Some men are playing a sport.	Entailment

Training data file Choose File no file selected

Test data file Choose File no file selected

Evaluate Dataset

Please choose a model to test in the following:

○ Bert ○ ESIM ○ fastText

Please upload your model prediction result on test data with the following format:

1. The data file should be in txt data format.

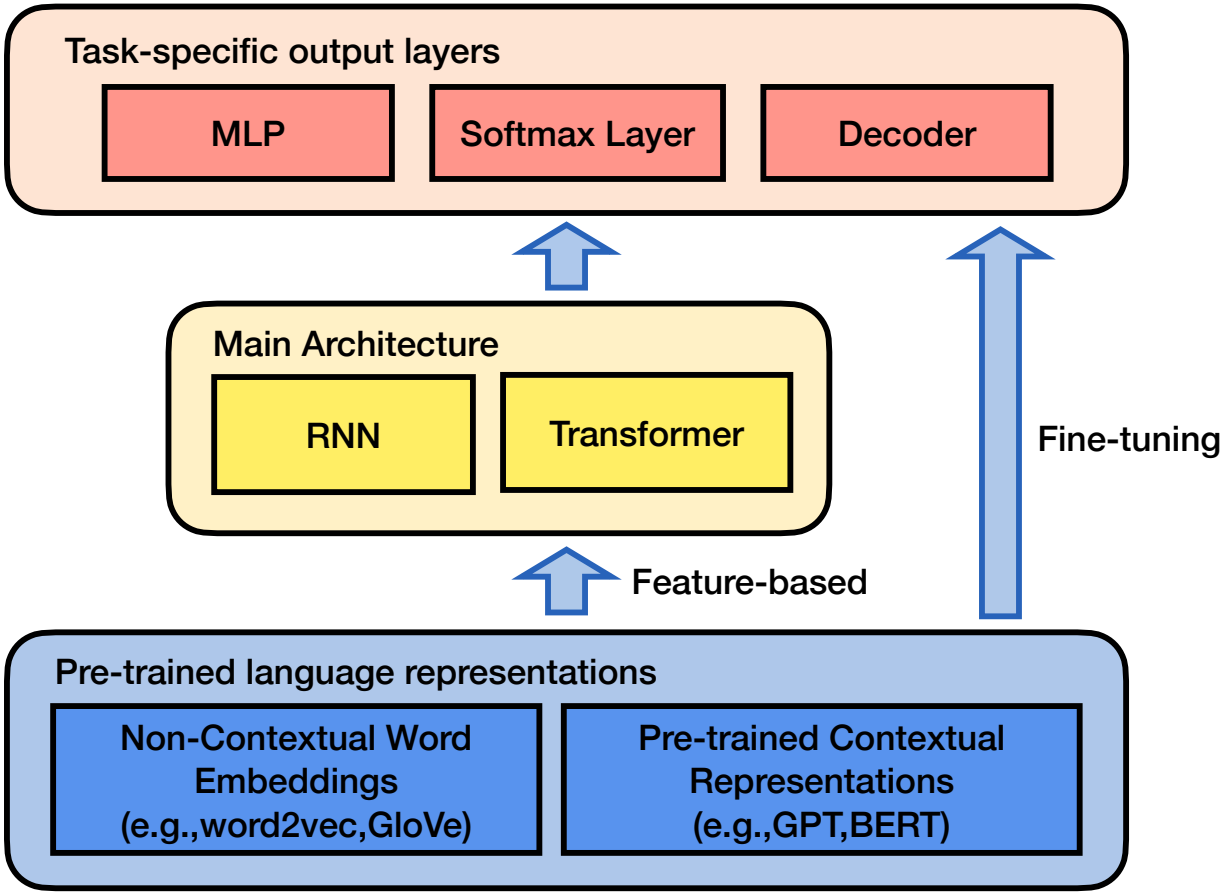
2. The data should contain four parts: ID, PREDICTION on each line and connected with “\t”.

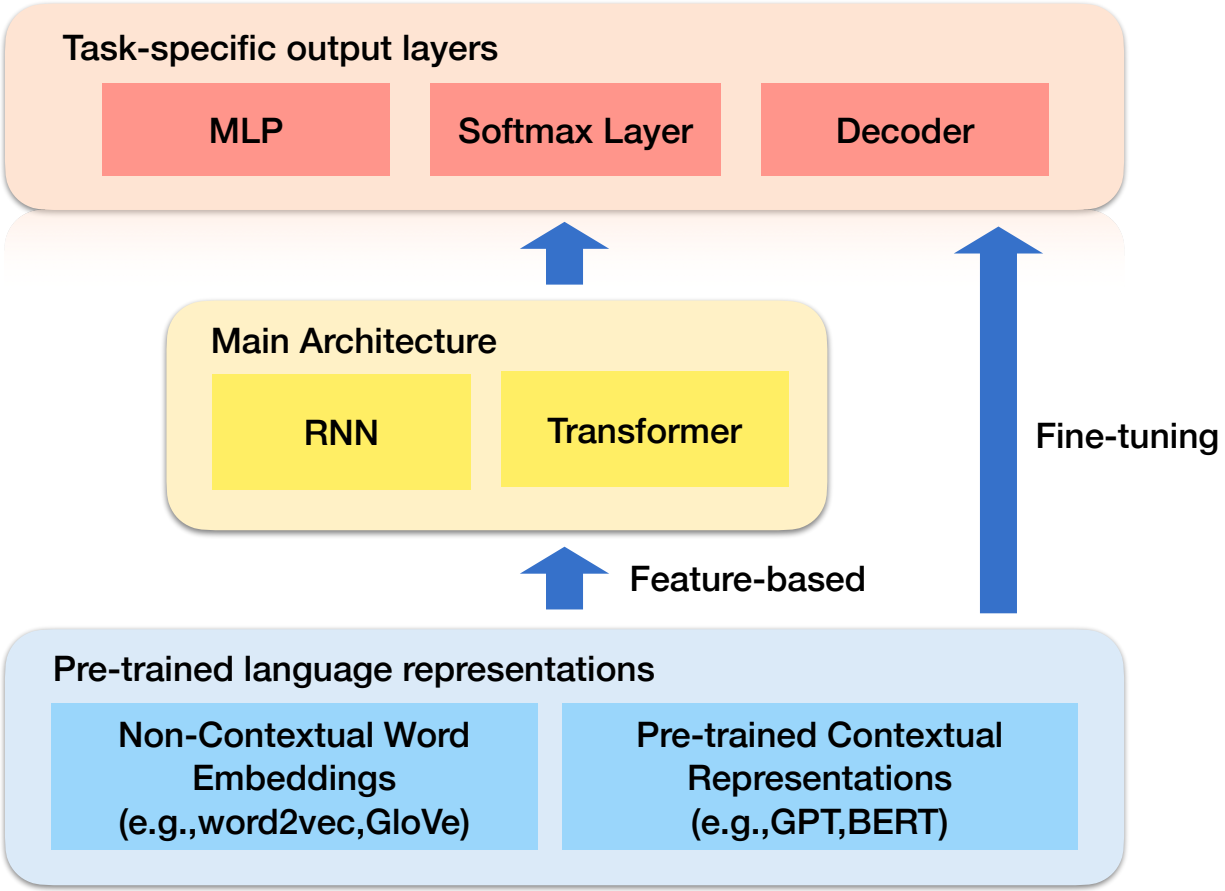
For example: 0\tentailement\n

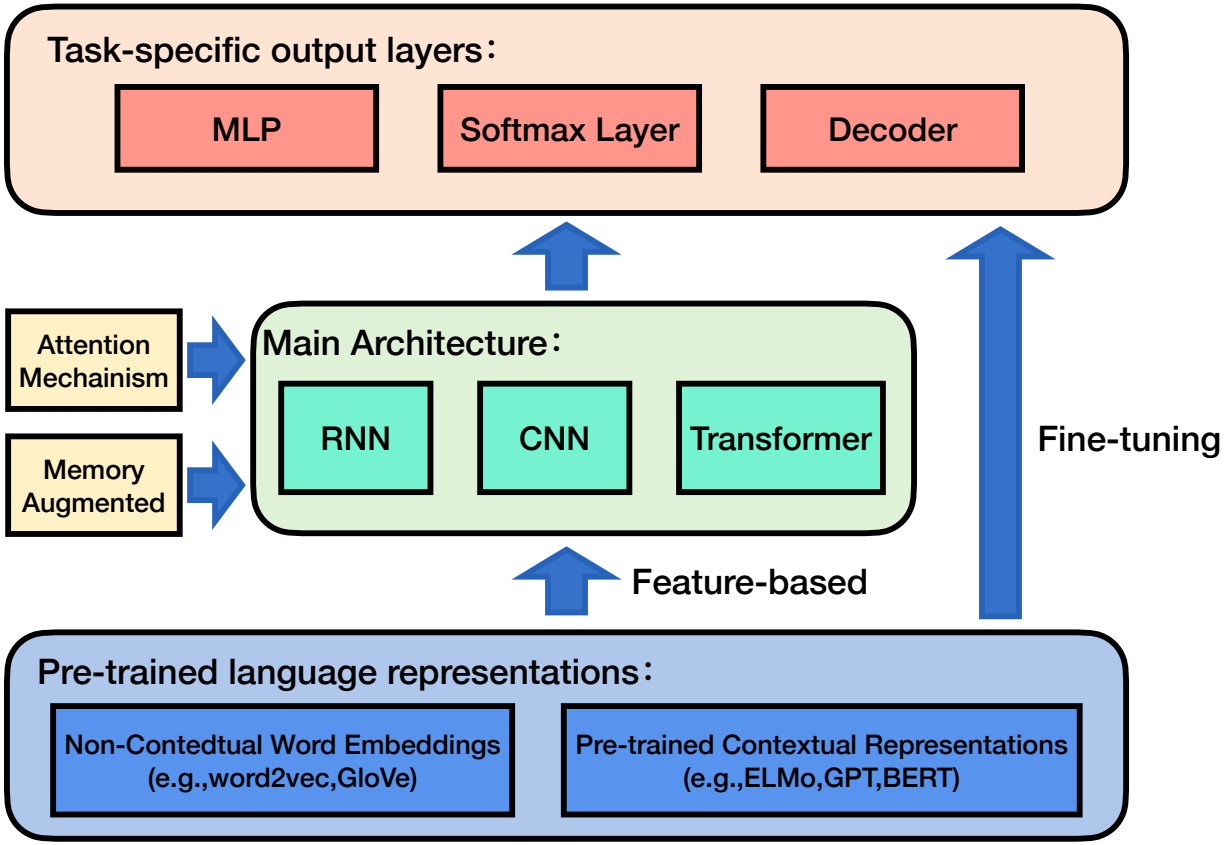
Test prediction file Choose File no file selected

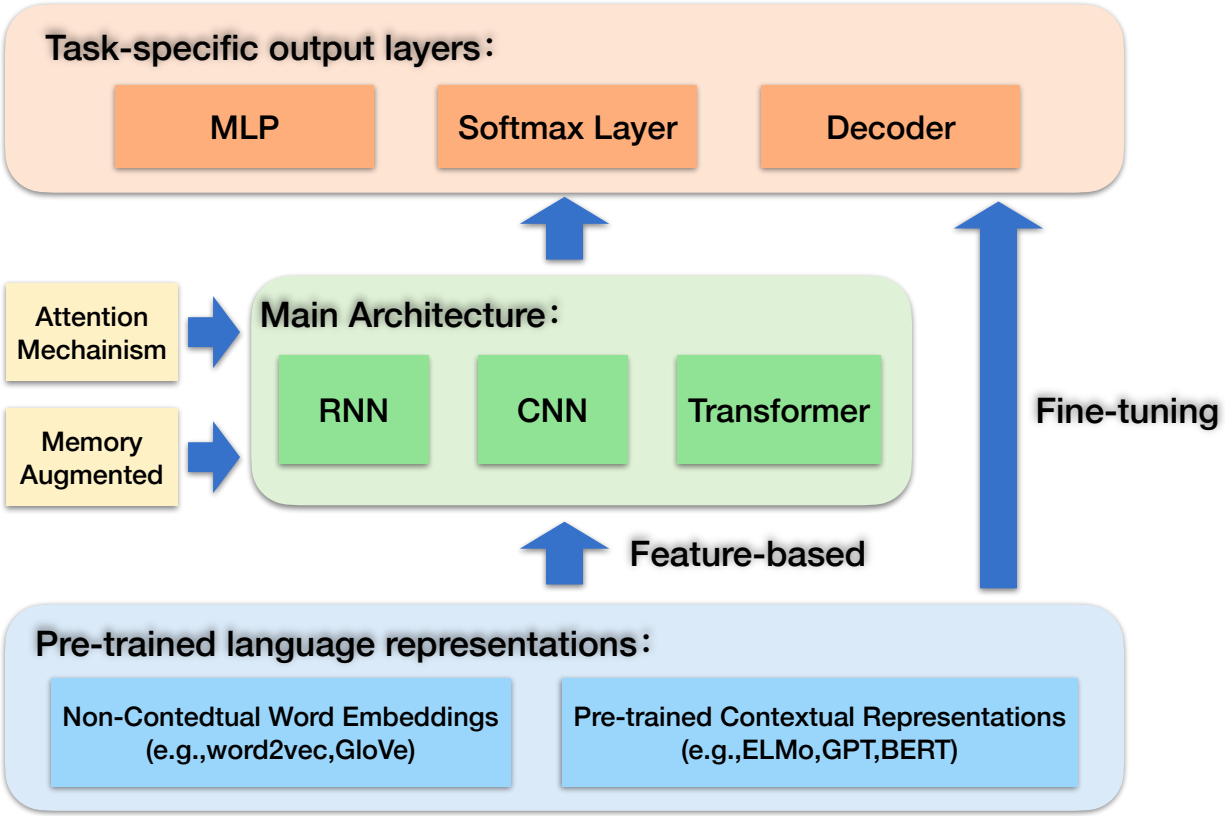
If you have test prediction file.

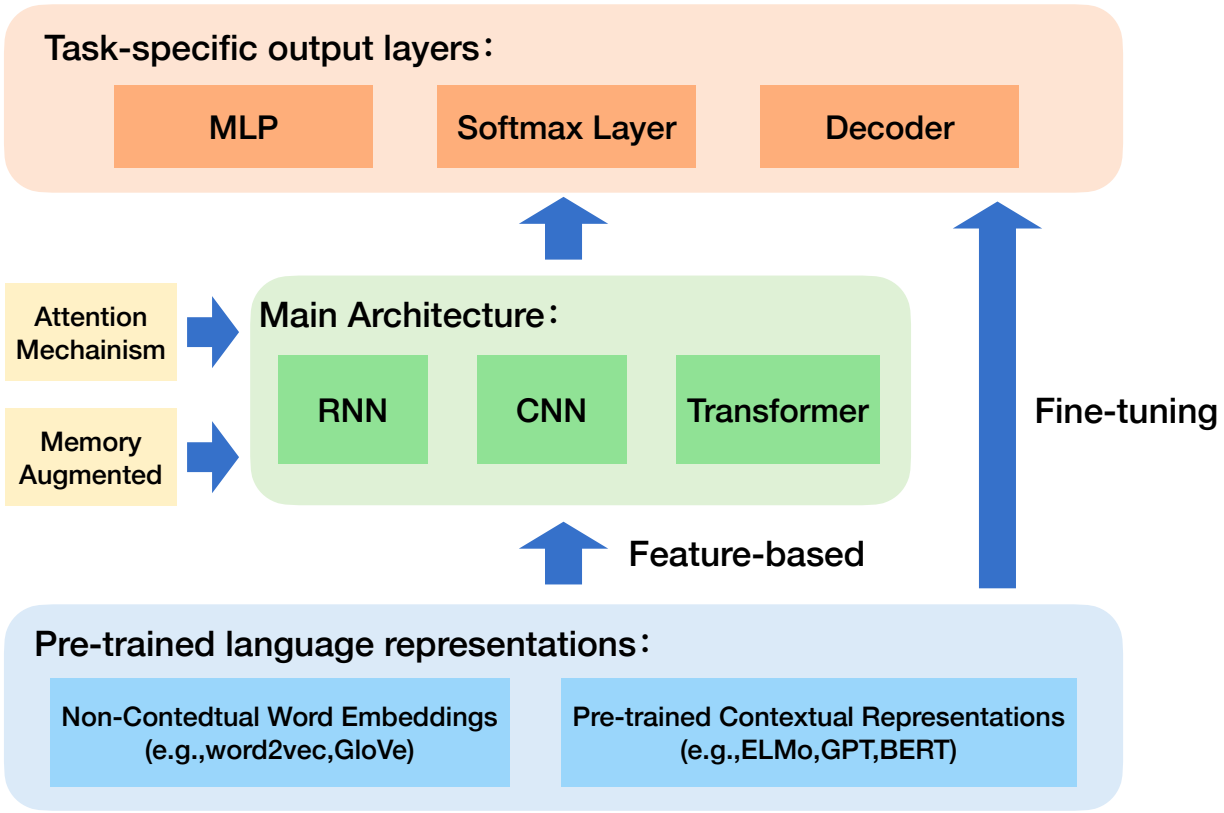
Evaluate Model

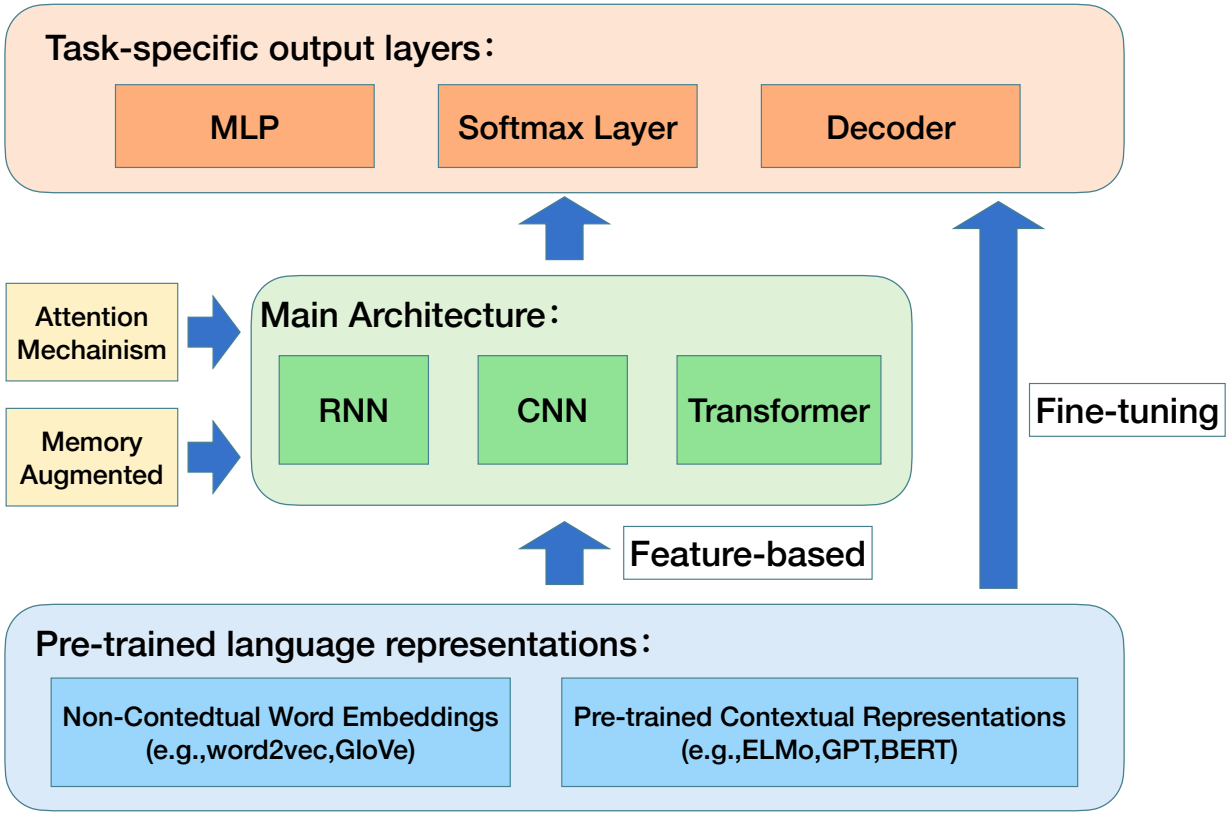

Task-specific output layers
MLP
Softmax Layer
Decoder
Main Architecture
RNN
Transformer
Feature-based
Fine-tuning
Pre-trained language representations
Non-Contextual Word Embeddings (e.g.,word2vec,GloVe)
Pre-trained Contextual Representations (e.g.,GPT,BERT)

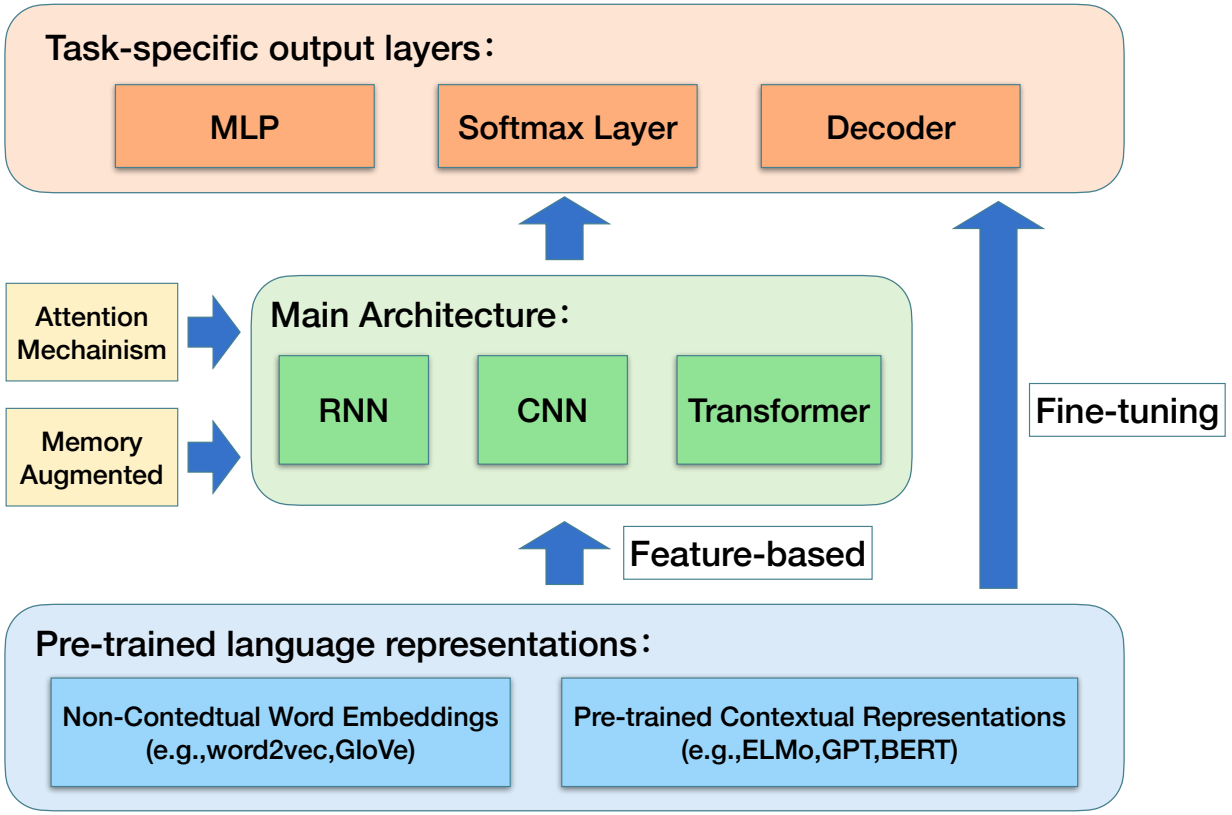

Task-specific output layers
MLP
Softmax Layer
Decoder
Main Architecture
RNN
Transformer
Feature-based
Fine-tuning
Pre-trained language representations
Non-Contextual Word Embeddings (e.g.,word2vec,GloVe)
Pre-trained Contextual Representations (e.g.,GPT,BERT)

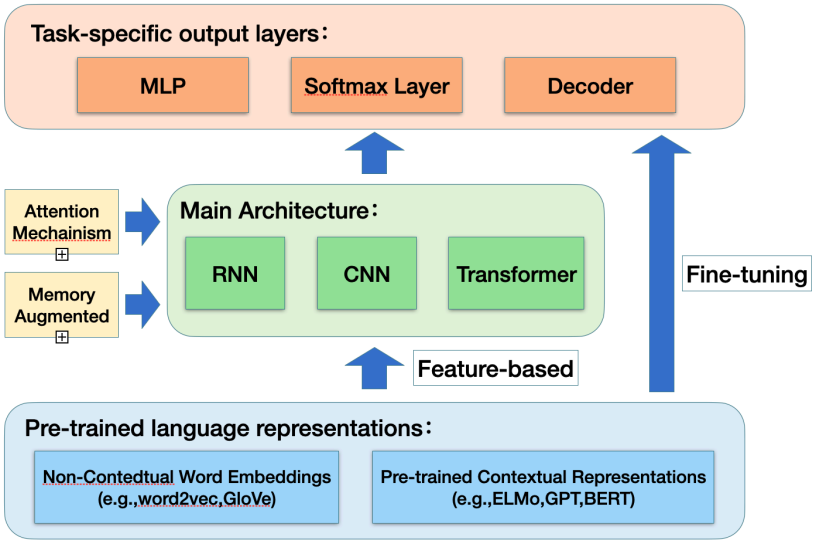

Task-specific output layers:
MLP
Softmax Layer
Decoder
Attention Mechainism
Memory Augmented
Main Architecture:
RNN
CNN
Transformer
Fine-tuning
Feature-based
Pre-trained language representations:
Non-Contedtual Word Embeddings (e.g.,word2vec,GloVe)
Pre-trained Contextual Representations (e.g.,ELMo,GPT,BERT)

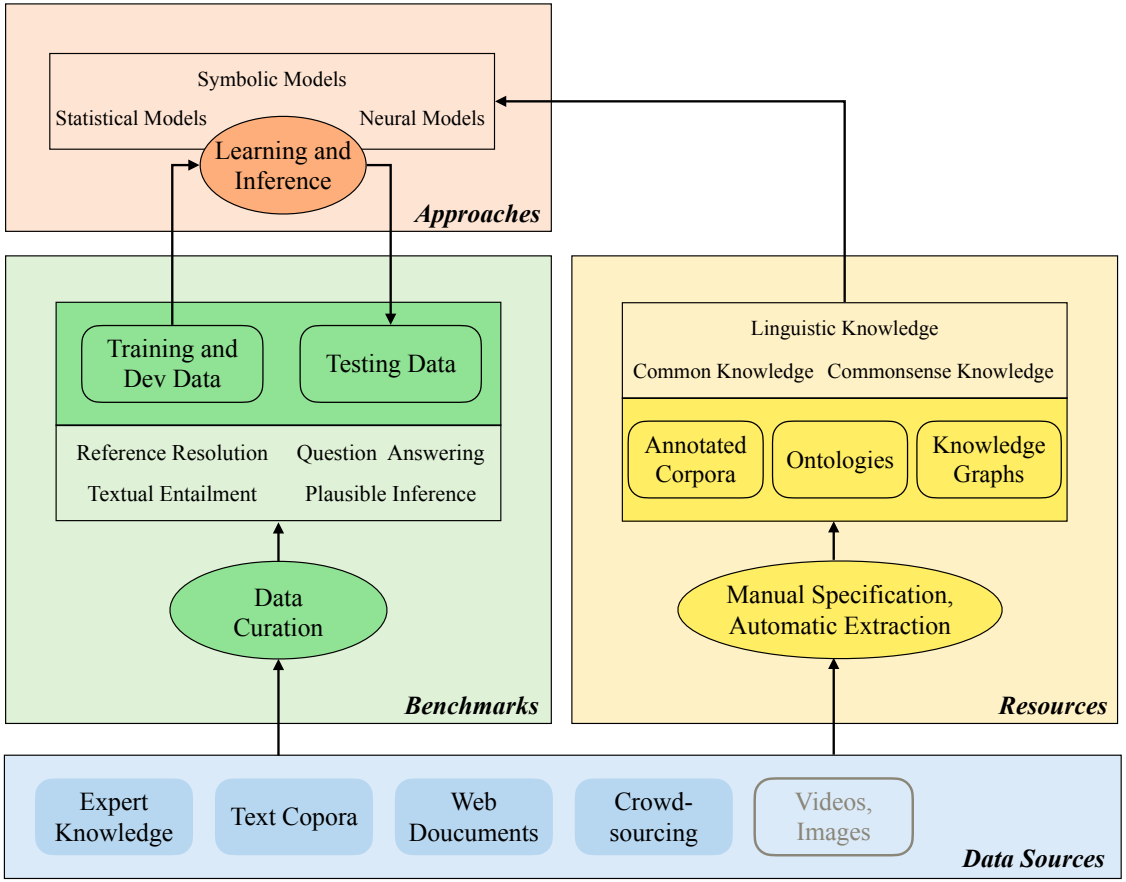

Task-specific output layers:
MLP
Softmax Layer
Decoder
Attention Mechainism
Memory Augmented
Main Architecture:
RNN
CNN
Transformer
Feature-based
Fine-tuning
Pre-trained language representations:
Non-Contedtual Word Embeddings (e.g.,word2vec,GloVe)
Pre-trained Contextual Representations (e.g.,ELMo,GPT,BERT)

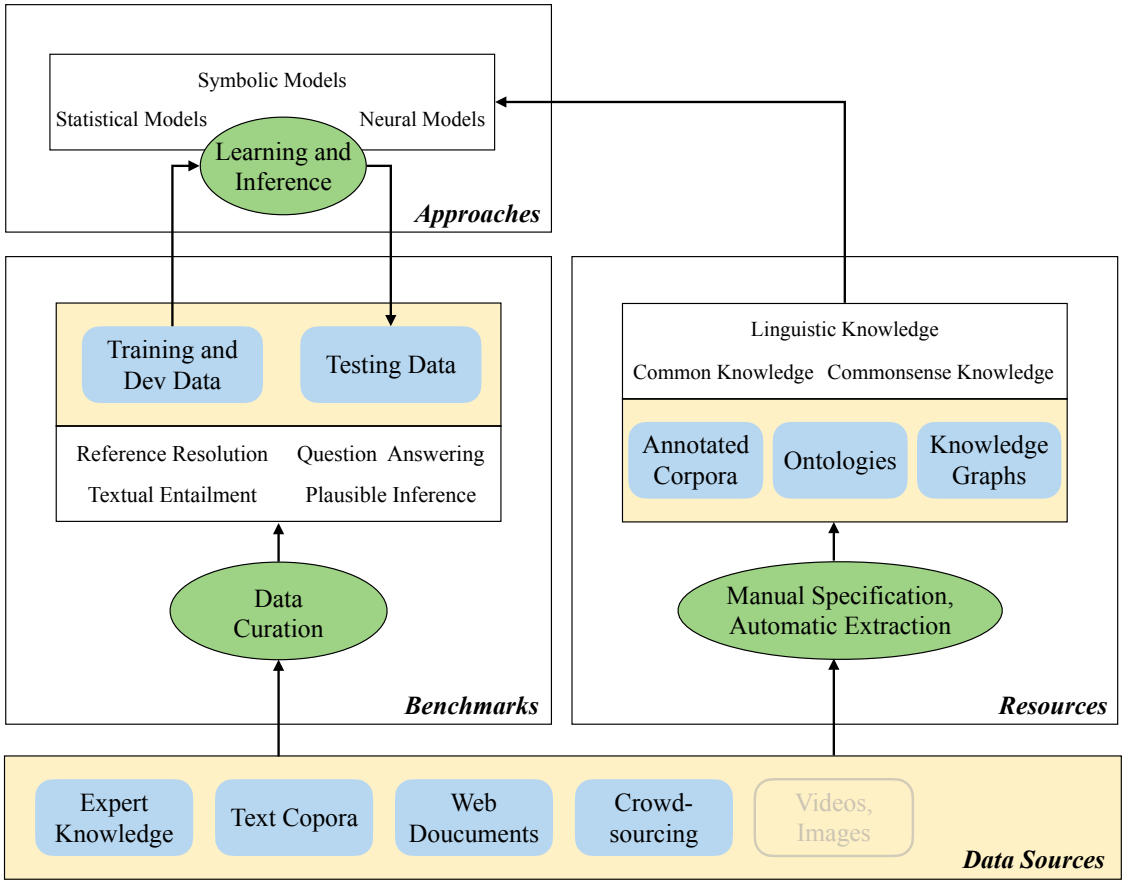

Task-specific output layers:
MLP
Softmax Layer
Decoder
Attention Mechainism
Memory Augmented
Main Architecture:
RNN
CNN
Transformer
Feature-based
Fine-tuning
Pre-trained language representations:
Non-Contedtual Word Embeddings (e.g.,word2vec,GloVe)
Pre-trained Contextual Representations (e.g.,ELMo,GPT,BERT)


Task-specific output layers:
MLP
Softmax Layer
Decoder
Attention Mechainism
Memory Augmented
Main Architecture:
RNN
CNN
Transformer
Fine-tuning
Feature-based
Pre-trained language representations:
Non-Contedtual Word Embeddings (e.g.,word2vec,GloVe)
Pre-trained Contextual Representations (e.g.,ELMo,GPT,BERT)


Task-specific output layers:
MLP
Softmax Layer
Decoder
Attention Mechainism
Memory Augmented
Main Architecture:
RNN
CNN
Transformer
Fine-tuning
Feature-based
Pre-trained language representations:
Non-Contedtual Word Embeddings (e.g.,word2vec,GloVe)
Pre-trained Contextual Representations (e.g.,ELMo,GPT,BERT)


Task-specific output layers:
MLP
Softmax Layer
Decoder
Attention Mechainism
Memory Augmented
Main Architecture:
RNN
CNN
Transformer
Fine-tuning
Feature-based
Pre-trained language representations:
Non-Contedtual Word Embeddings (e.g.,word2vec,GloVe)
Pre-trained Contextual Representations (e.g.,ELMo,GPT,BERT)

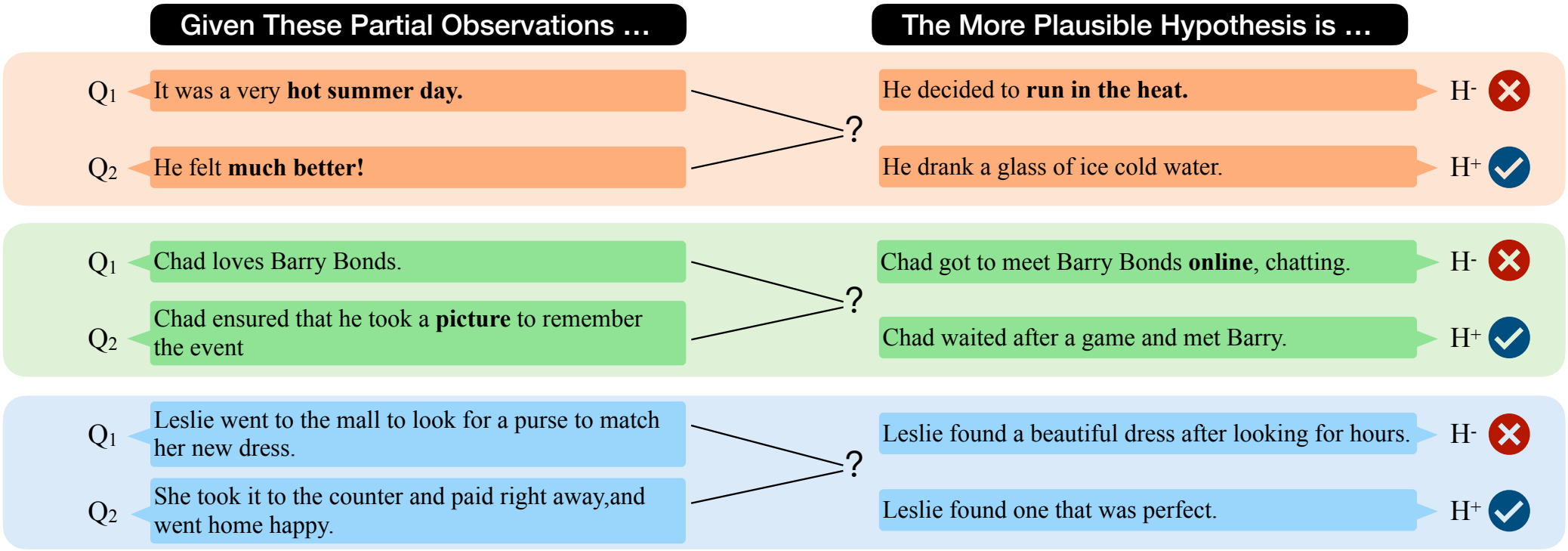

Symbolic Models
Statistical Models
Neural Models
Learning and Inference
Approaches
Training and Dev Data
Testing Data
Reference Resolution
Question Answering
Textual Entailment
Plausible Inference
Data Curation
Benchmarks
Linguistic Knowledge
Common Knowledge
Commonsense Knowledge
Annotated Corpora
Ontologies
Knowledge Graphs
Manual Specification, Automatic Extraction
Resources
Expert Knowledge
Text Copora
Web Doucuments
Crowd-sourcing
Videos, Images
Data Sources

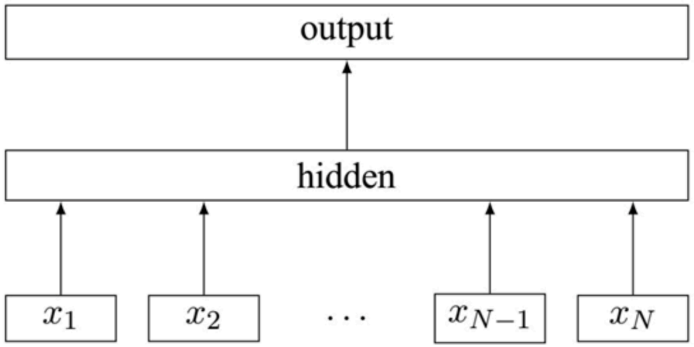

Symbolic Models
Statistical Models
Neural Models
Learning and Inference
Approaches
Training and Dev Data
Testing Data
Reference Resolution
Question Answering
Textual Entailment
Plausible Inference
Data Curation
Benchmarks
Linguistic Knowledge
Common Knowledge
Commonsense Knowledge
Annotated Corpora
Ontologies
Knowledge Graphs
Manual Specification, Automatic Extraction
Resources
Expert Knowledge
Text Copora
Web Doucuments
Crowd-sourcing
Videos, Images
Data Sources

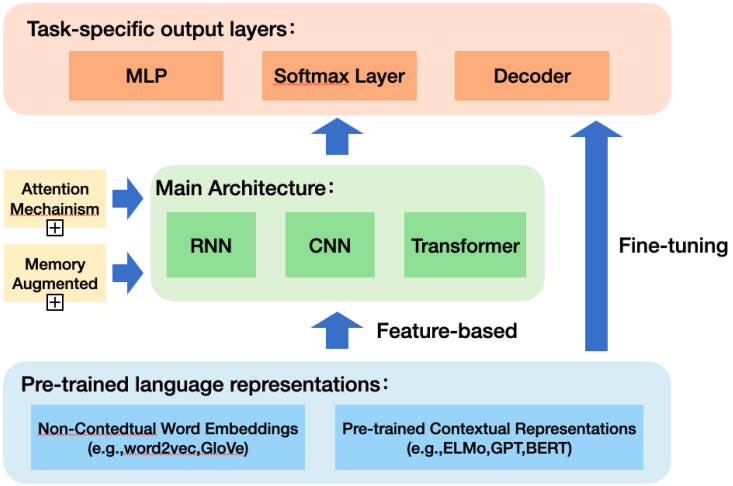
Premise: A man pulling items on a cart.

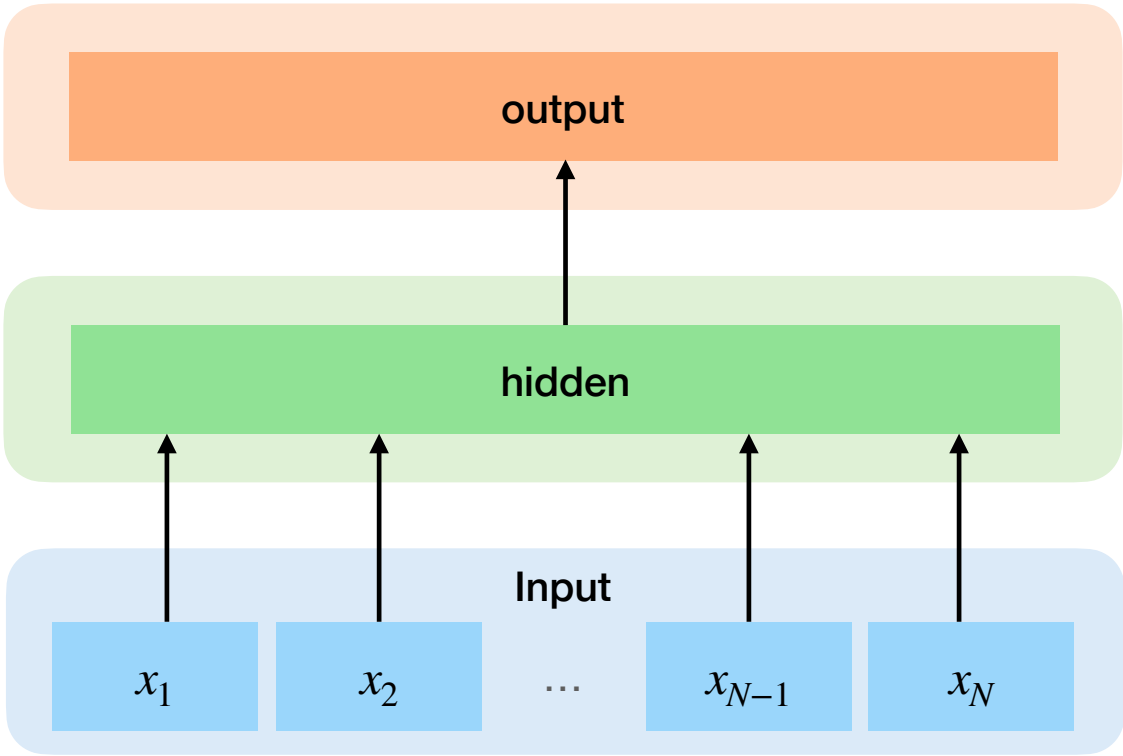
Hypothesis: A man is pushing a baby carriage. ✓ Contradiction

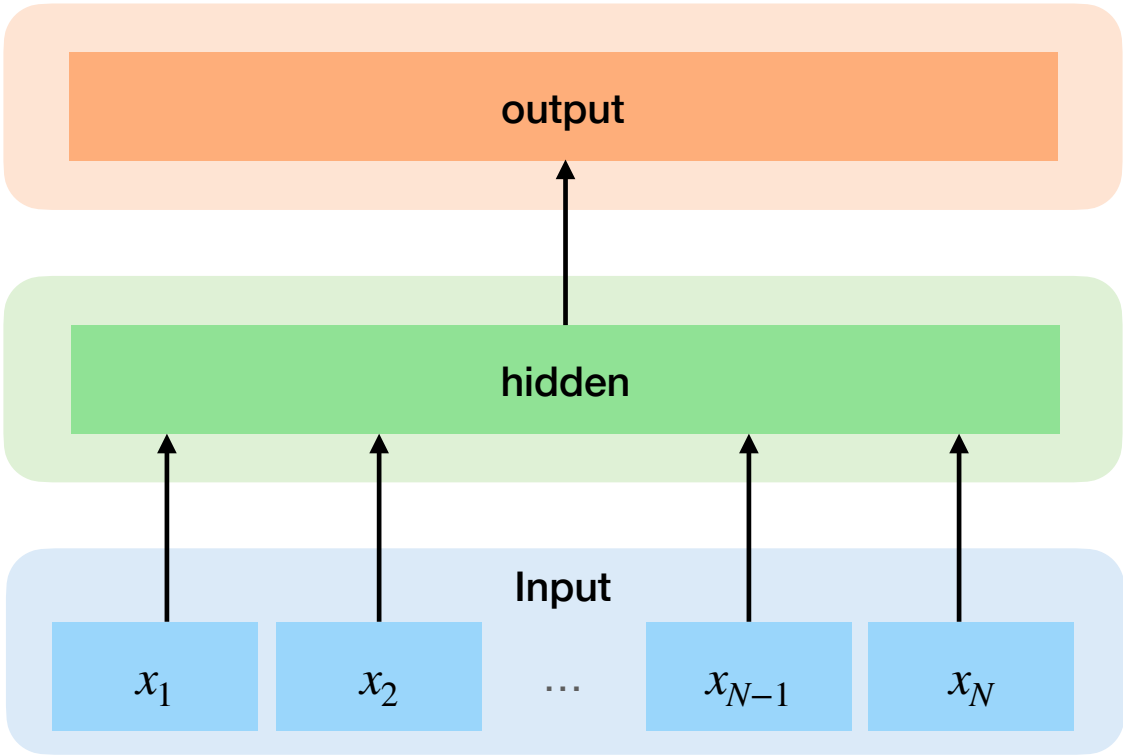
Hypothesis': A man is carring a baby carriage. ✗ Entailment


Given These Partial Observations ...
The More Plausible Hypothesis is ...
Q1
It was a very hot summer day.
Q2
He felt much better!
?
He decided to run in the heat.
H-
He drank a glass of ice cold water.
H+
Q1
Chad loves Barry Bonds.
Q2
Chad ensured that he took a picture to remember the event
?
Chad got to meet Barry Bonds online, chatting.
H-
Chad waited after a game and met Barry.
H+
Q1
Leslie went to the mall to look for a purse to match her new dress.
Q2
She took it to the counter and paid right away,and went home happy.
?
Leslie found a beautiful dress after looking for hours.
H-
Leslie found one that was perfect.
H+


output
hidden
x_1
x_2
$\cdots$
x_{N-1}
x_N


Task-specific output layers:
MLP
Softmax Layer
Decoder
Attention Mechanism
Memory Augmented
Main Architecture:
RNN
CNN
Transformer
Feature-based
Fine-tuning
Pre-trained language representations:
Non-Contedtual Word Embeddings (e.g.,word2vec,GloVe)
Pre-trained Contextual Representations (e.g.,ELMo,GPT,BERT)


output
hidden
Input
x_1
x_2
$\cdots$
x_{N-1}
x_N


output
hidden
Input
x_1
x_2
...
x_{N-1}
x_N

文段（Passage）：The role of teacher is often formal and ongoing, carried out at a school or other place of formal education. In many countries, a person who wishes to become a teacher must first obtain specified professional qualifications or credentials from a university or college. These professional qualifications may include the study of pedagogy, the science of teaching. Teachers, like other professionals, may have to continue their education after they qualify, a process known as continuing professional development. Teachers may use a lesson plan to facilitate student learning, providing a course of study which is called the curriculum.

问题（Question）：What can a teacher use to help students learn?

答案（Answer）：lesson plan

前提 (Premise) : A man pulling items on a cart.

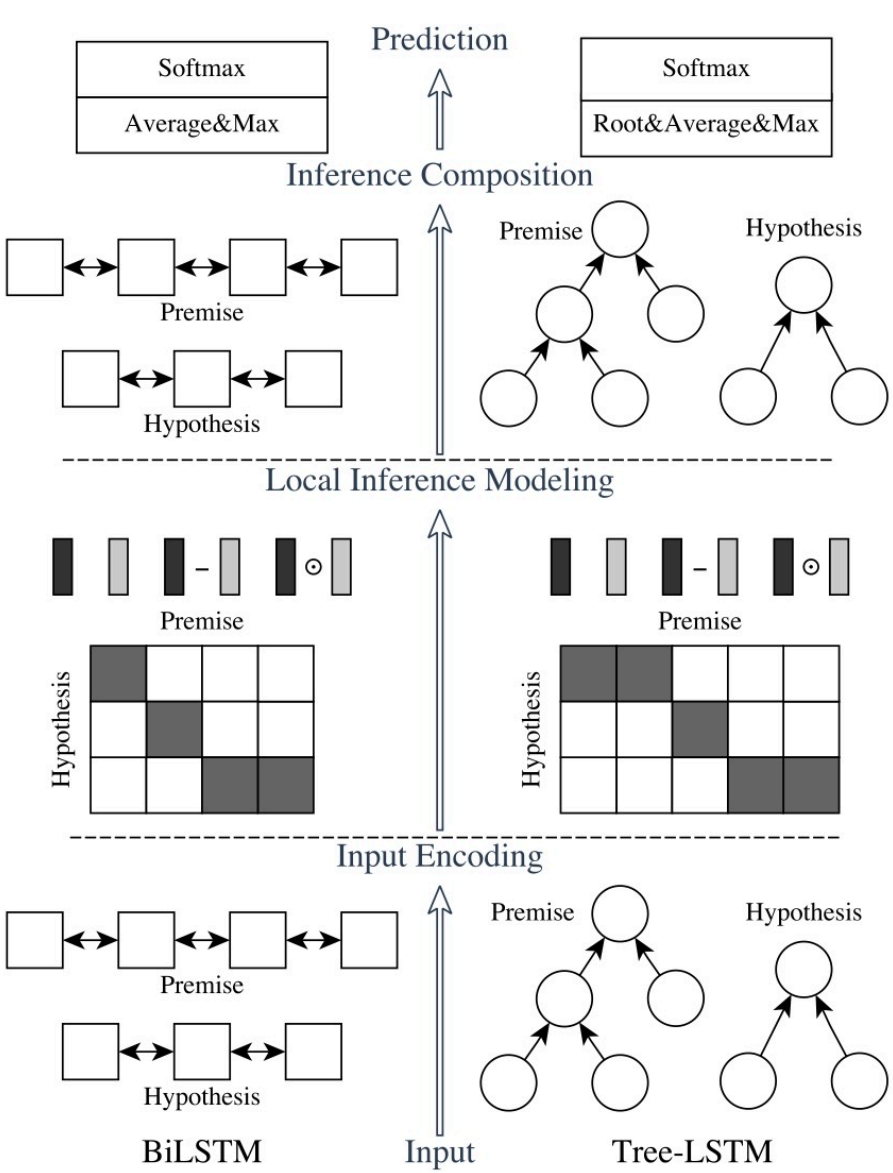
假设 (Hypothesis) : A man is pushing a baby carriage.

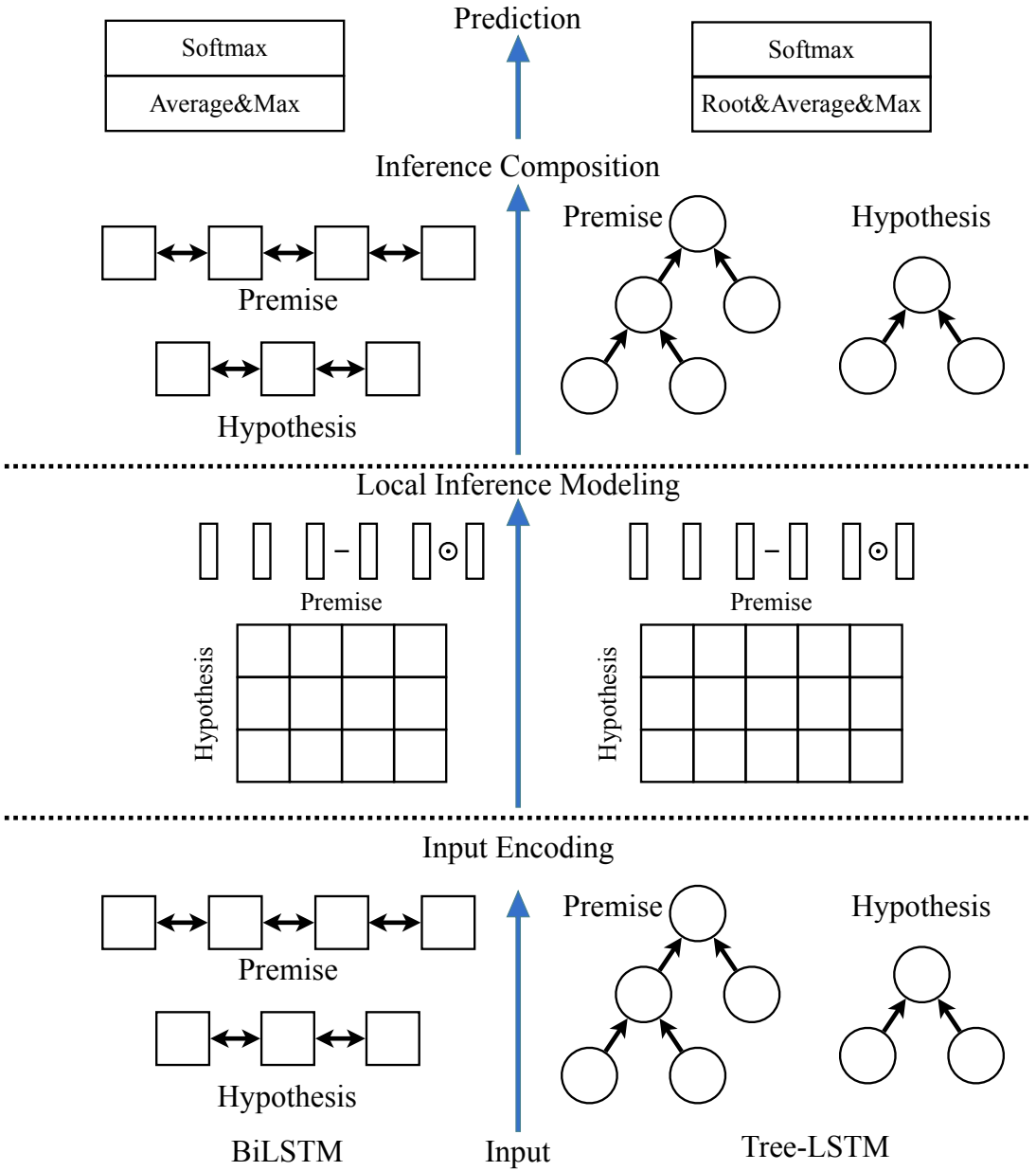
选择 (Choice) : Entailment, Contradiction, Neutral

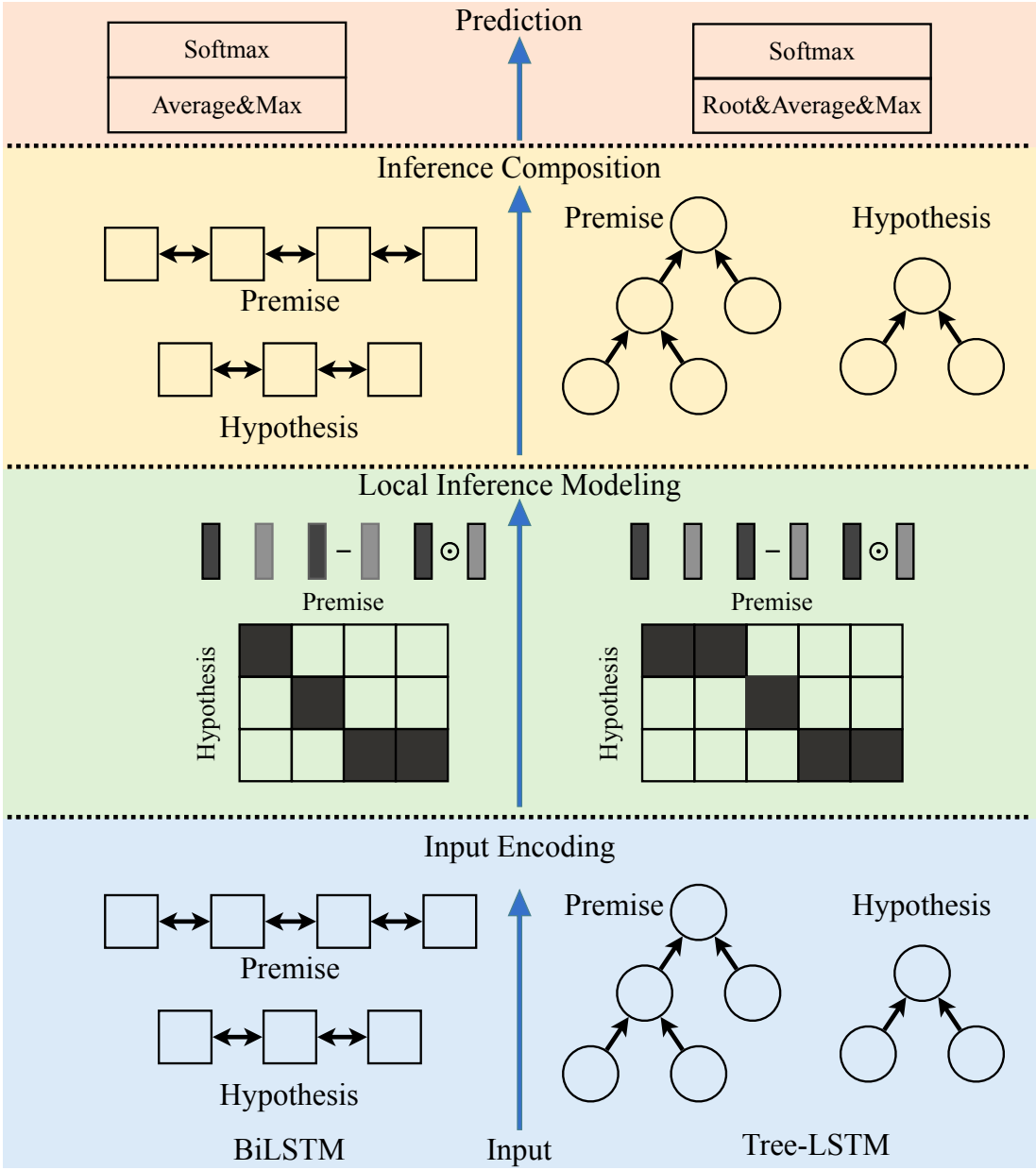
前提 (Premise) :_A man pulling items on a cart.

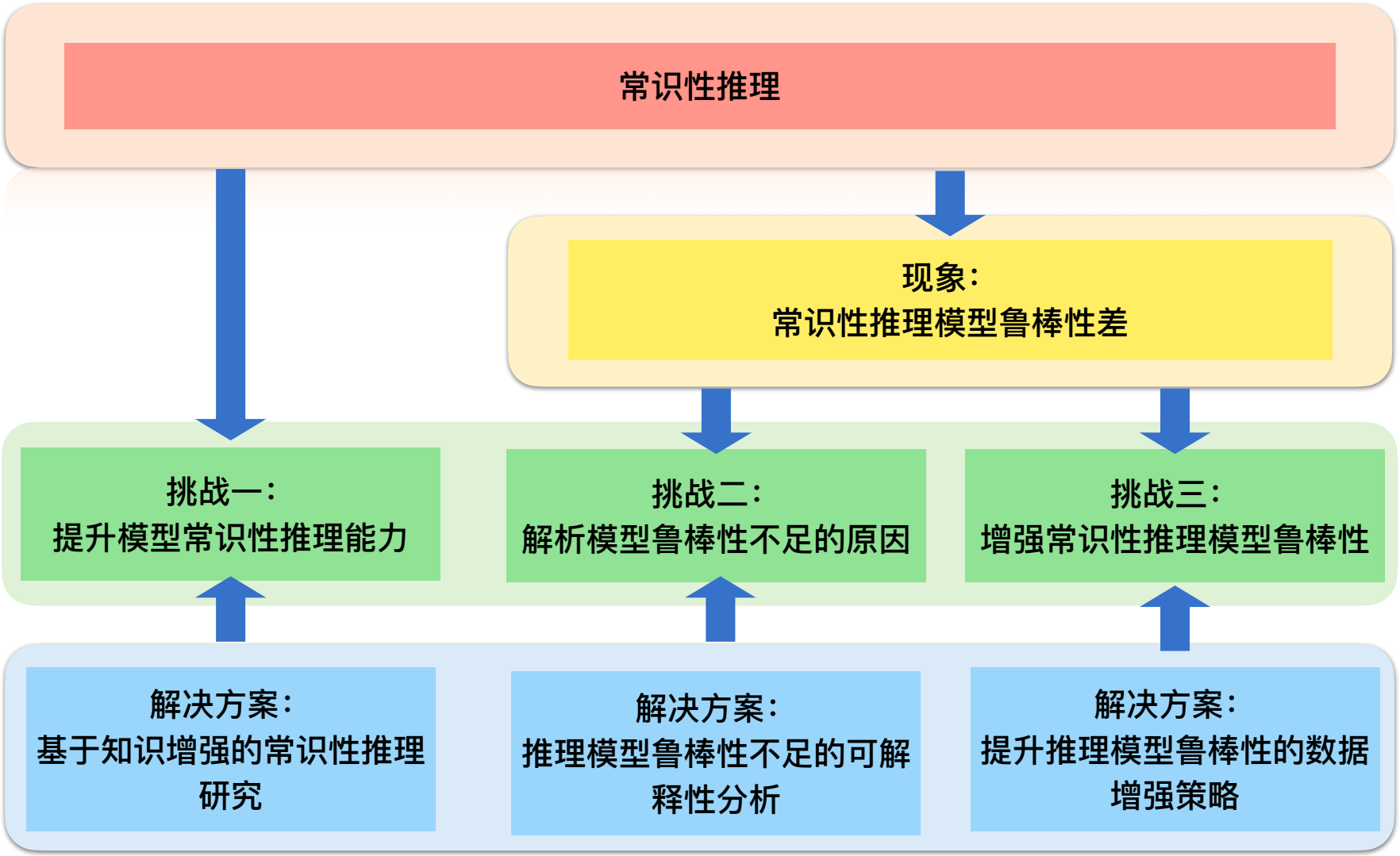
假设 (Hypothesis) : A man is pushing a baby carriage.

选择 (Choice) : Entailment, Contradiction, Neutral


Softmax
Average&Max
Prediction
Softmax
Root&Average&Max
Inference Composition
Premise
Hypothesis
Premise
Hypothesis
Local Inference Modeling
Premise
Premise
Hypothesis
Hypothesis
Input Encoding
Premise
Hypothesis
Premise
Hypothesis
BiLSTM
Input
Tree-LSTM


Softmax
Average&Max
Prediction
Softmax
Root&Average&Max
Inference Composition
Premise
Hypothesis
Premise
Hypothesis
Local Inference Modeling
Premise
Premise
Hypothesis
Hypothesis
Input Encoding
Premise
Hypothesis
Premise
Hypothesis
BiLSTM
Input
Tree-LSTM


Prediction
Softmax
Average&Max
Softmax
Root&Average&Max
Inference Composition
Premise
Hypothesis
Premise
Hypothesis
Local Inference Modeling
Premise
Hypothesis
Premise
Hypothesis
Input Encoding
Premise
Hypothesis
Premise
Hypothesis
BiLSTM
Input
Tree-LSTM


常识性推理
现象：
常识性推理模型鲁棒性差
挑战一：
提升模型常识性推理能力
挑战二：
解析模型鲁棒性不足的原因
挑战三：
增强常识性推理模型鲁棒性
解决方案：
基于知识增强的常识性推理研究
解决方案：
推理模型鲁棒性不足的可解释性分析
解决方案：
提升推理模型鲁棒性的数据增强策略