You can lead a horse to water….: Representing vs. Using Features in Neural NLP

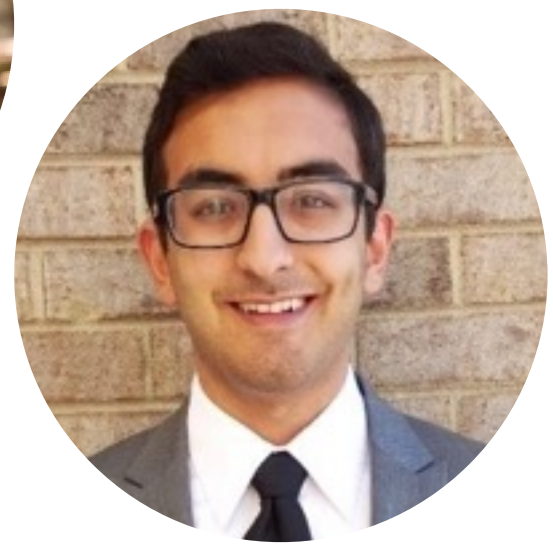
Ellie Pavlick
Department of Computer Science
Brown University

BROWN

Google Research

Shout out to my many coauthors!

Ian
Tenney



Amil
Merchant

Elahe
Rahimtoroghi

Dipanjan
Das

Charlie
Lovering

Rohan
Jha

Tal
Linzen

Tom
McCoy

Past ~2 years:
What do deep LMs know about language?

Past ~2 years:
What do deep LMs know about language?

What types of features
representations encode?
(“Probing Classifiers”)

Past ~2 years:
What do deep LMs know about language?

What types of features **representations** encode?
("Probing Classifiers")

Do models **behave** like they are using these features?
("Challenge Tasks")

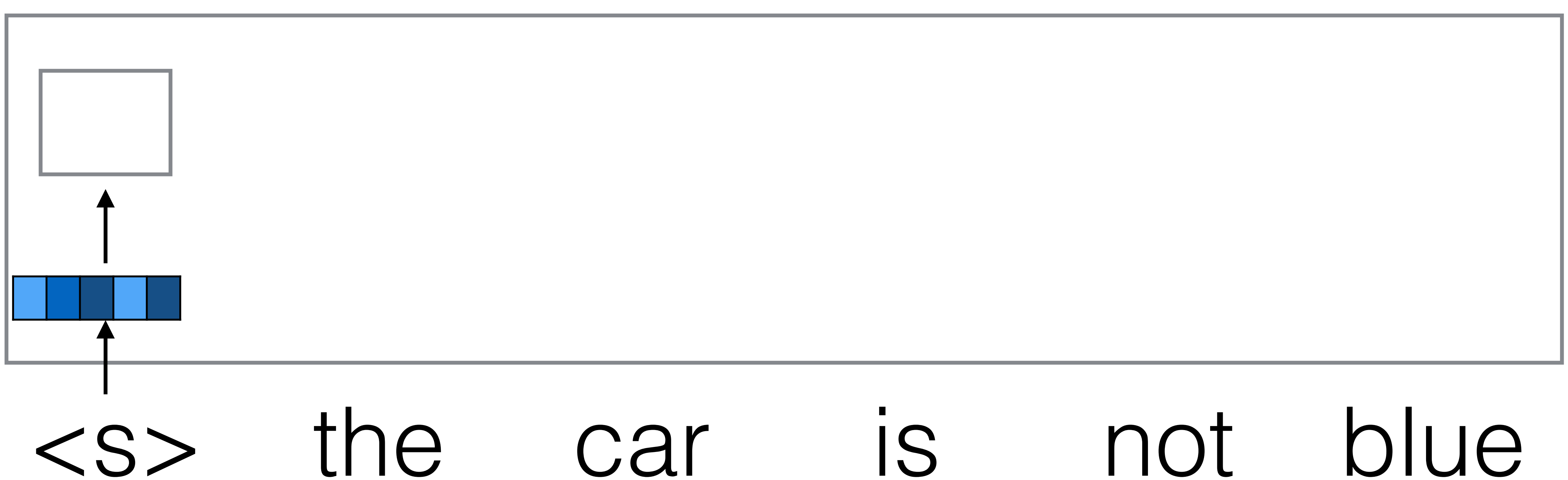
Past ~2 years:
What do deep LMs know about language?

What types of features **representations** encode?
("Probing Classifiers")

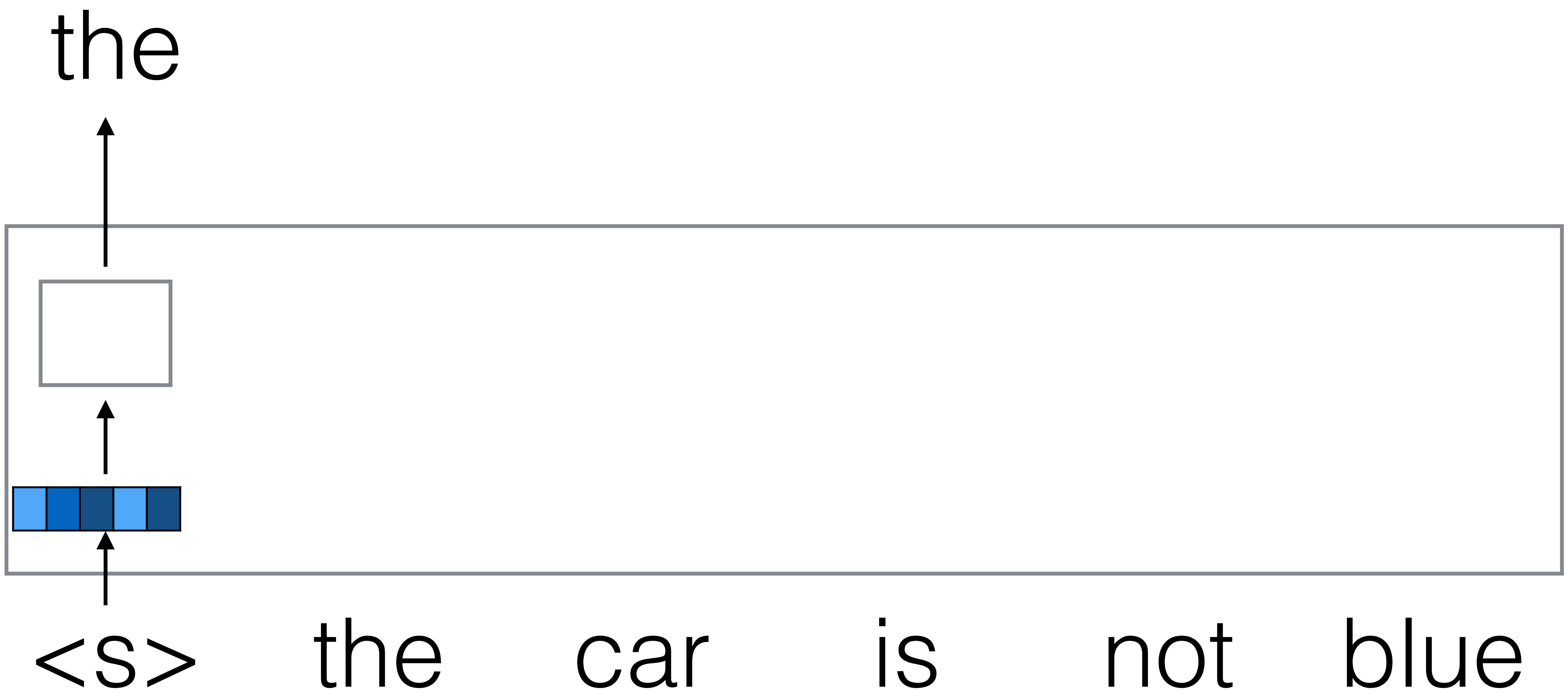
Do models **behave** like they are using these features?
("Challenge Tasks")

Is such-and-such feature encoded by the representation?

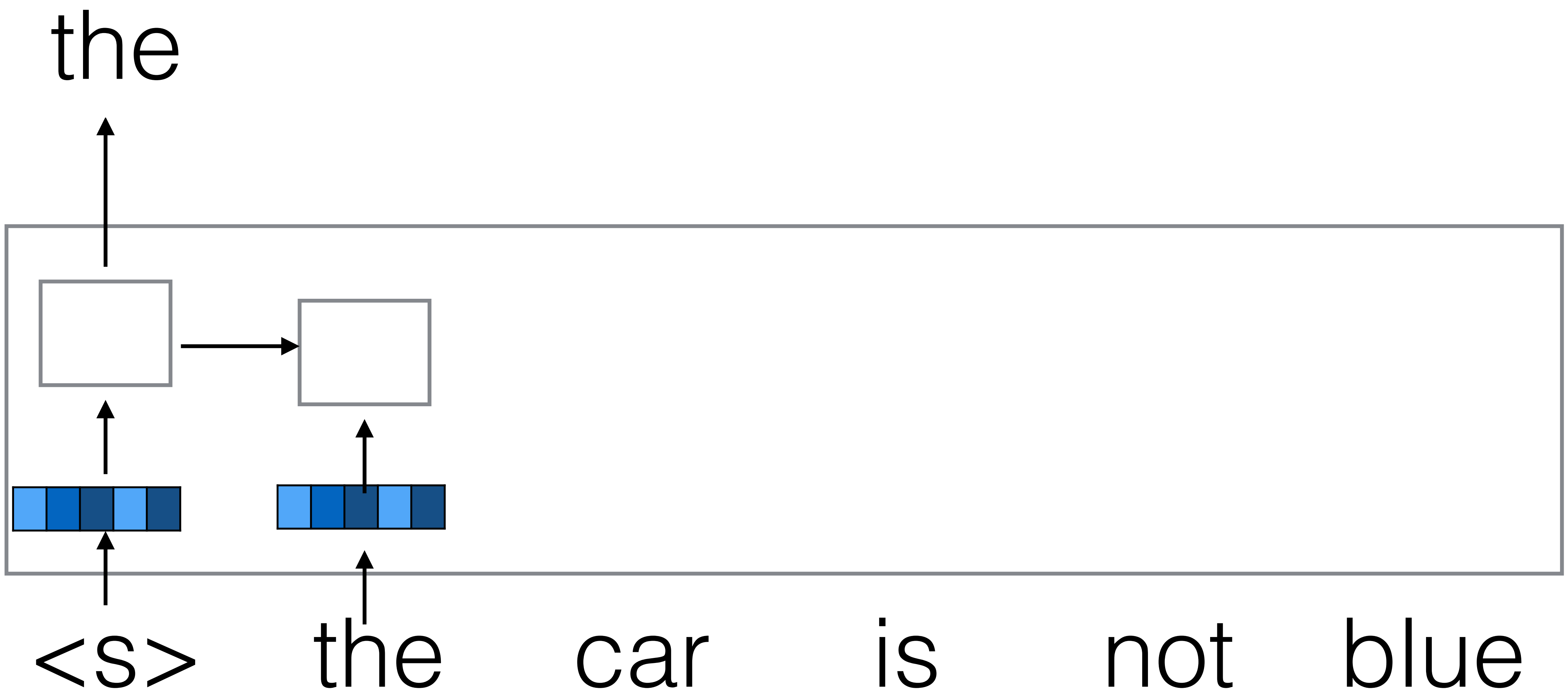
Is such-and-such feature encoded by the representation?



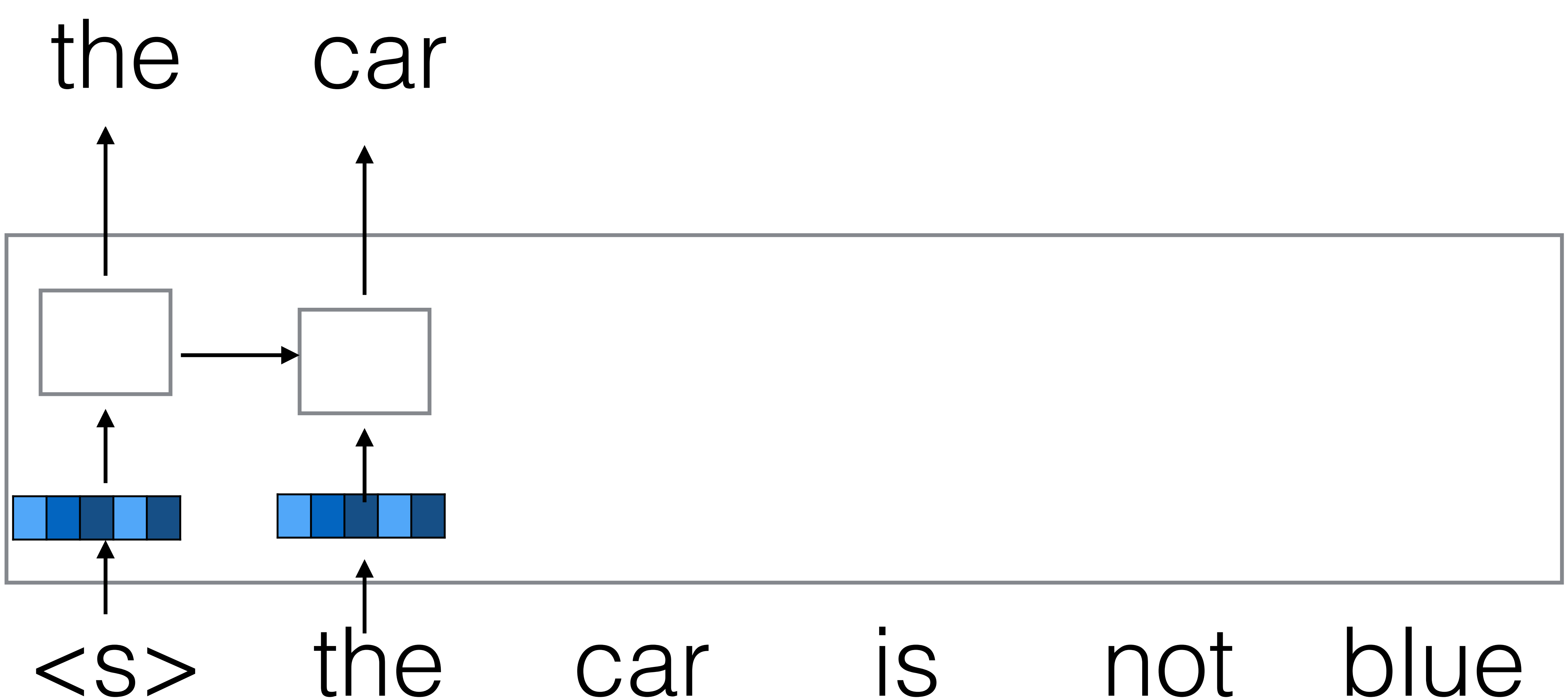
Is such-and-such feature encoded by the representation?



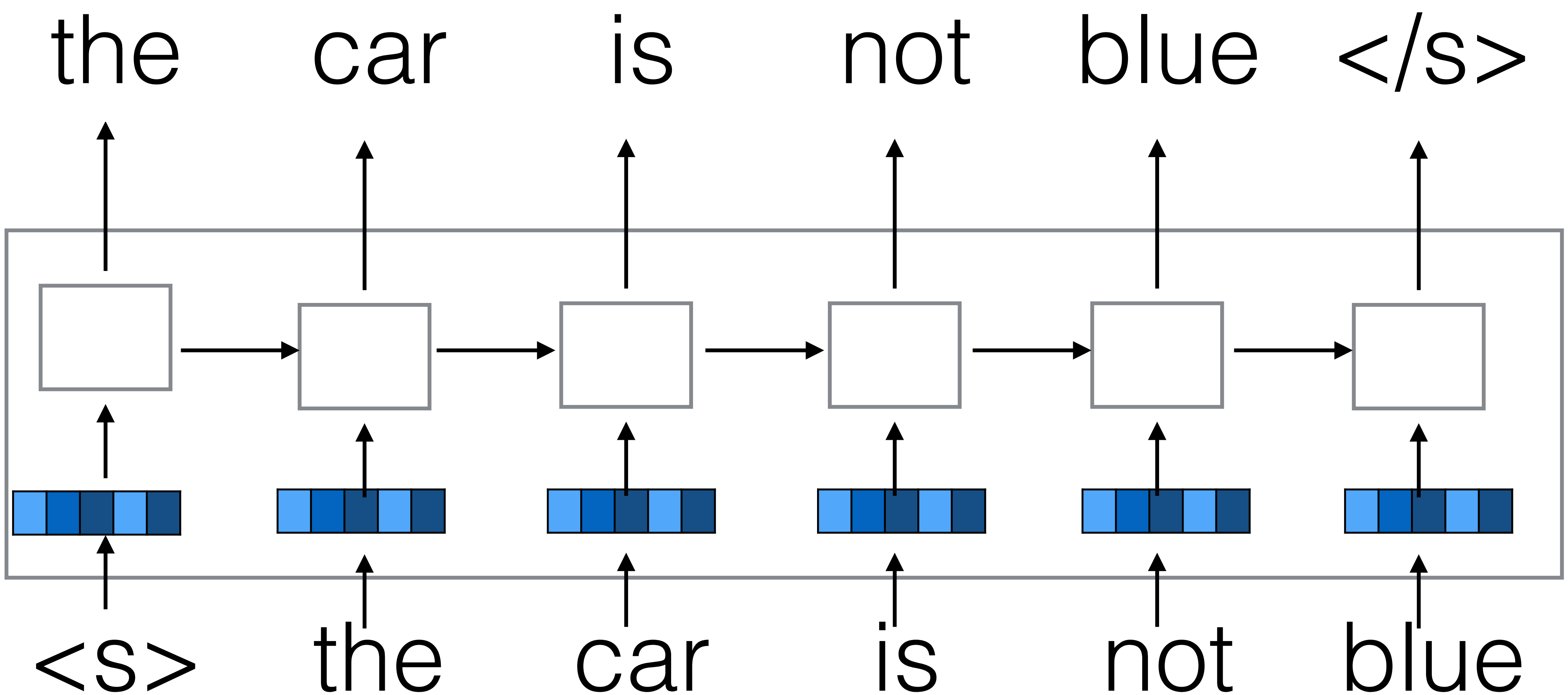
Is such-and-such feature encoded by the representation?



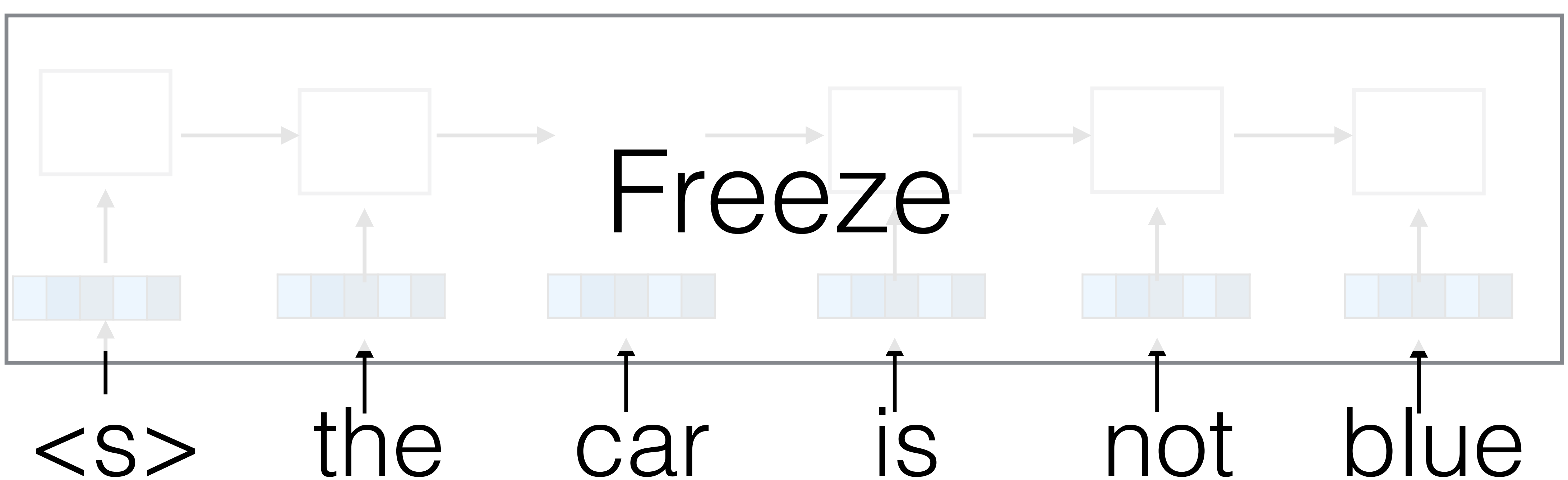
Is such-and-such feature encoded by the representation?



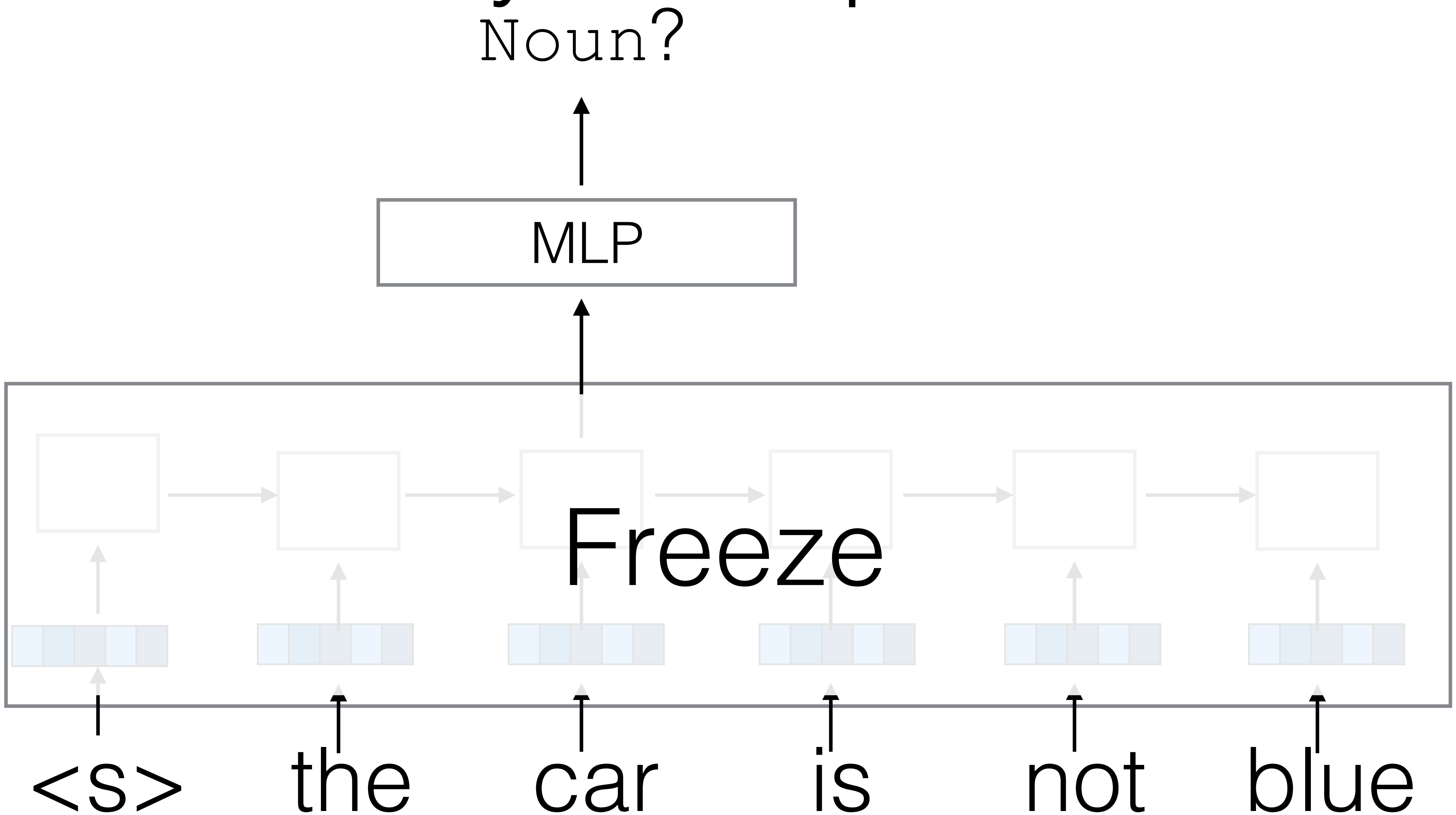
Is such-and-such feature encoded by the representation?



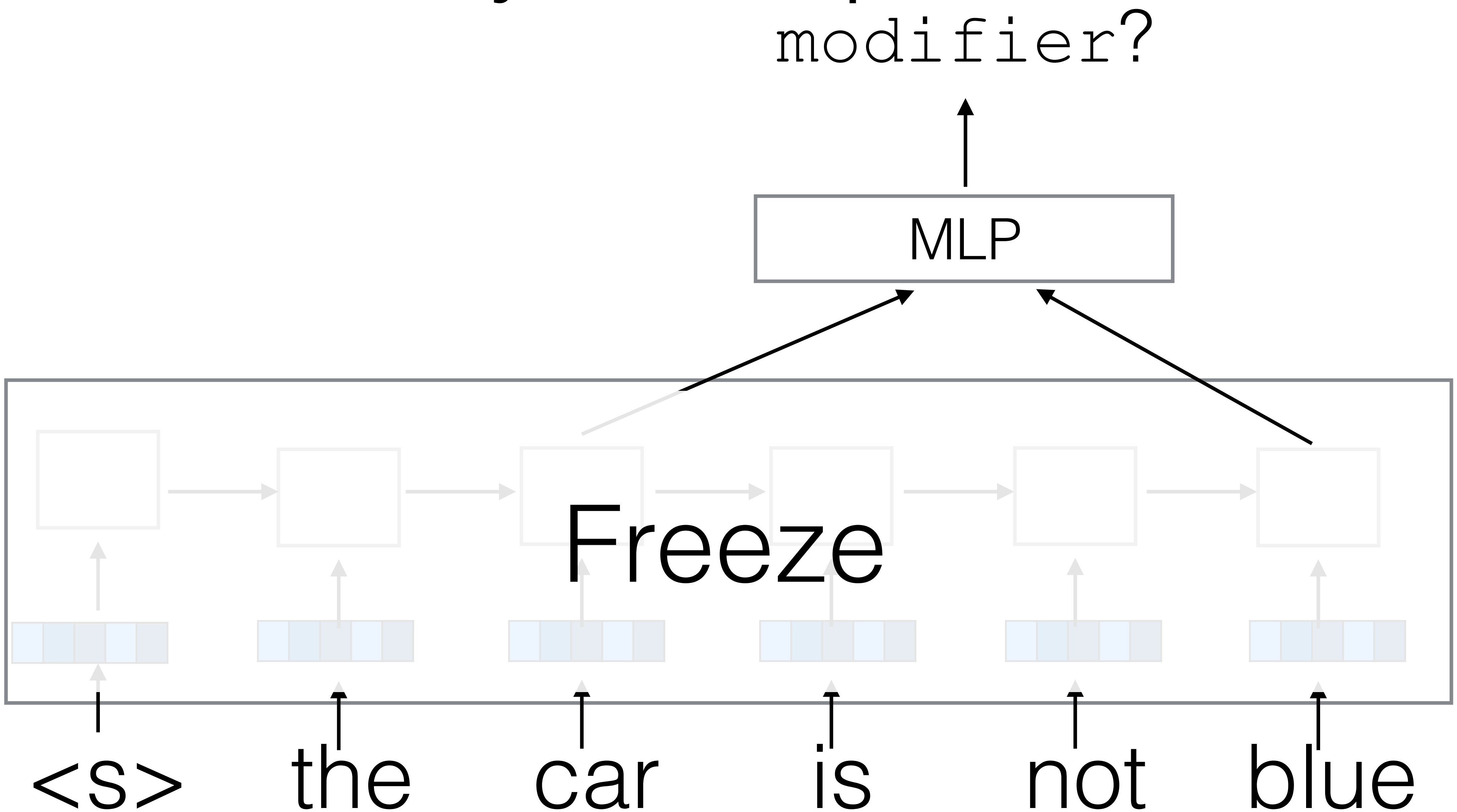
Is such-and-such feature encoded by the representation?



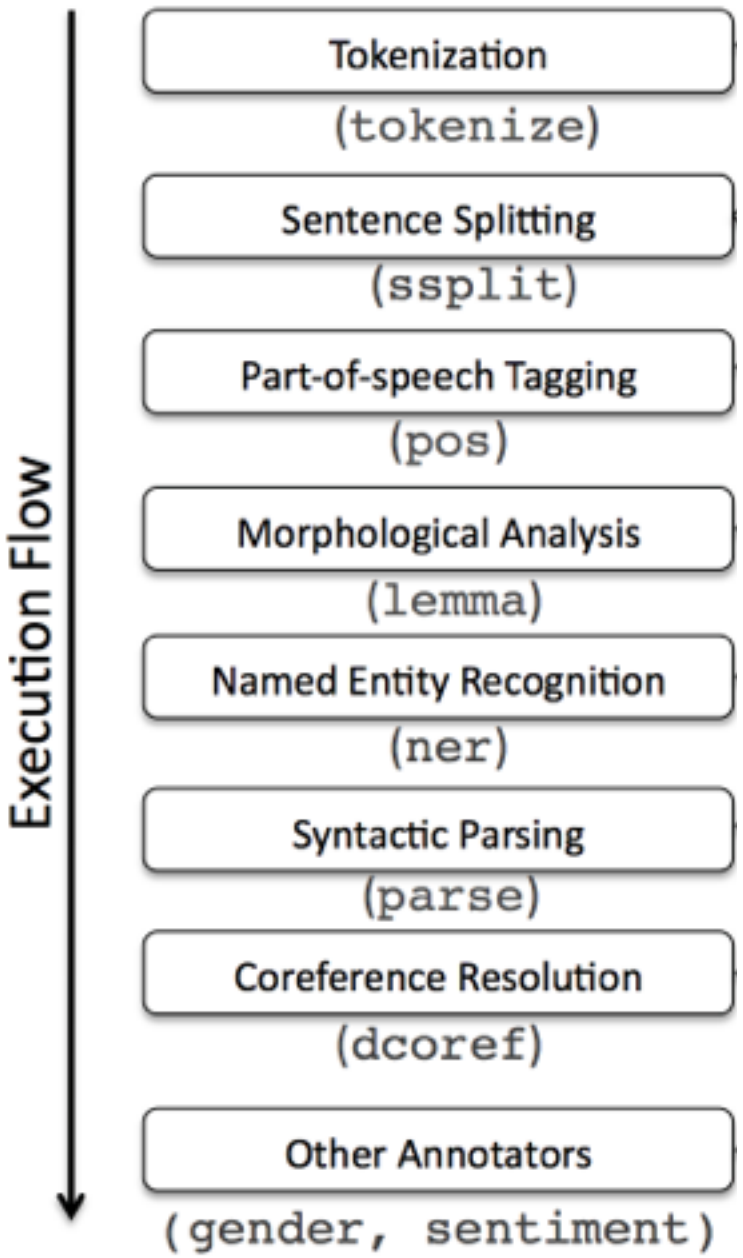
Is such-and-such feature encoded by the representation?



Is such-and-such feature encoded by the representation?



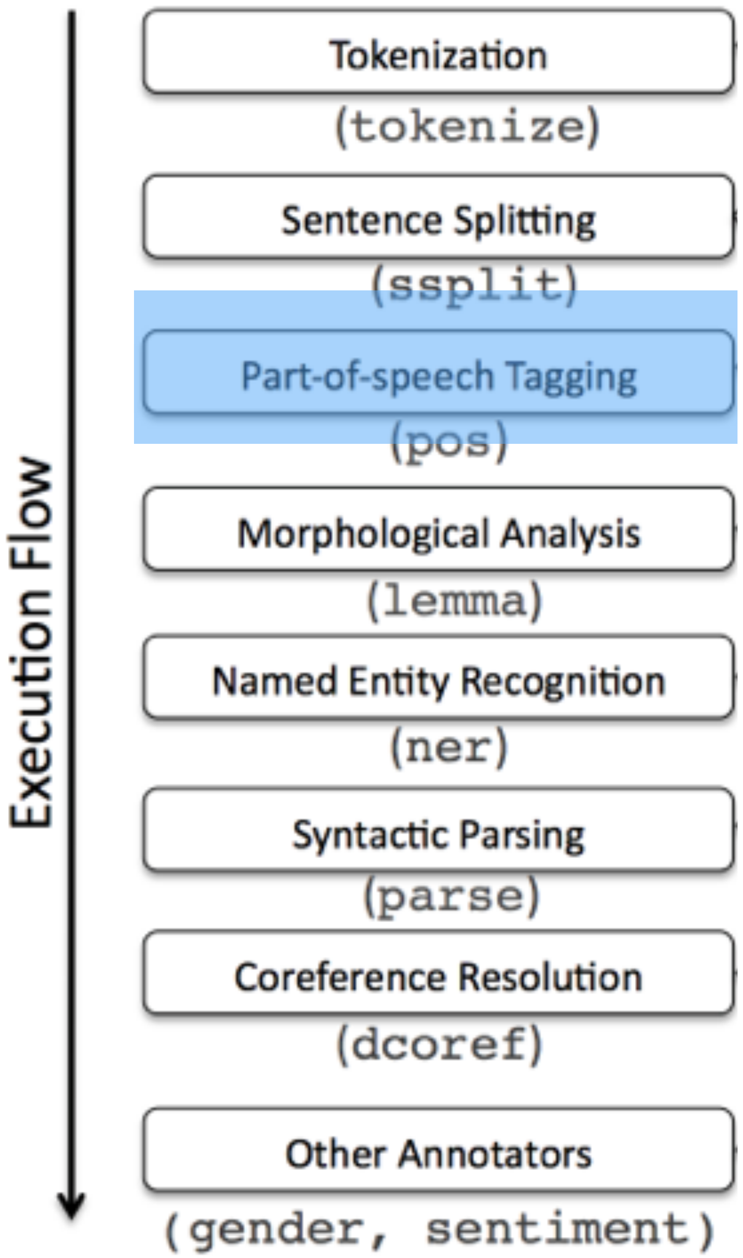
Is such-and-such feature encoded by the representation?



The important thing about Disney is that it is a global brand.

Probing Sentence Structure in Contextualized Word Representations. Tenney et al. (ICLR 2019)

Is such-and-such feature encoded by the representation?

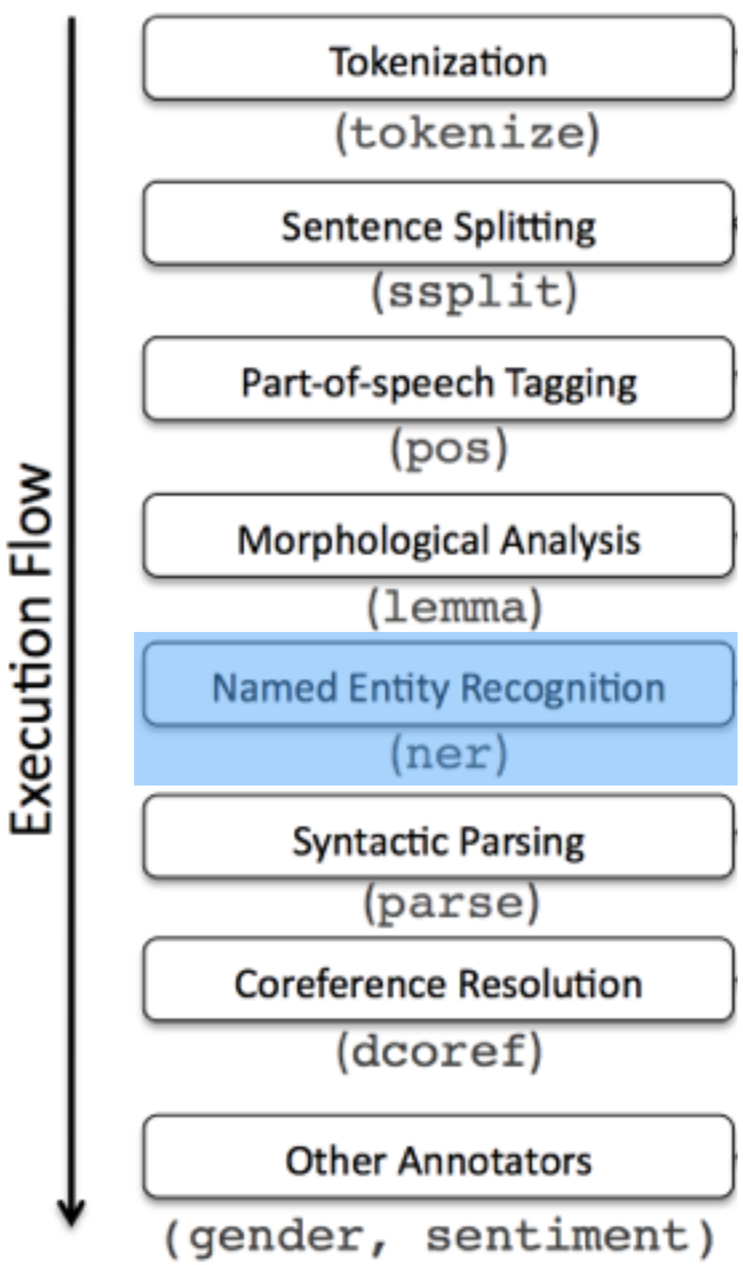


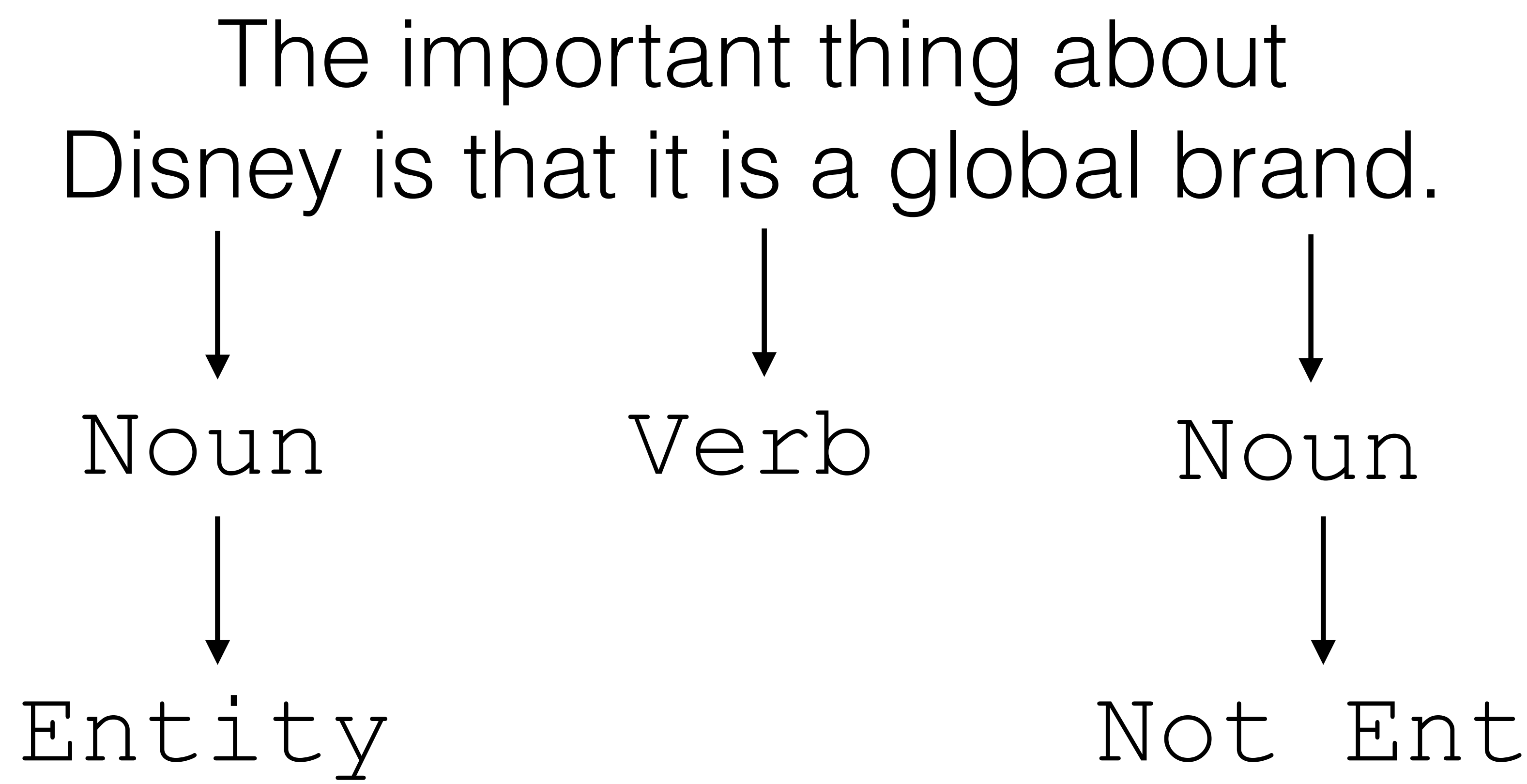
The important thing about Disney is that it is a global brand.

Noun Verb Noun

Probing Sentence Structure in Contextualized Word Representations. Tenney et al. (ICLR 2019)

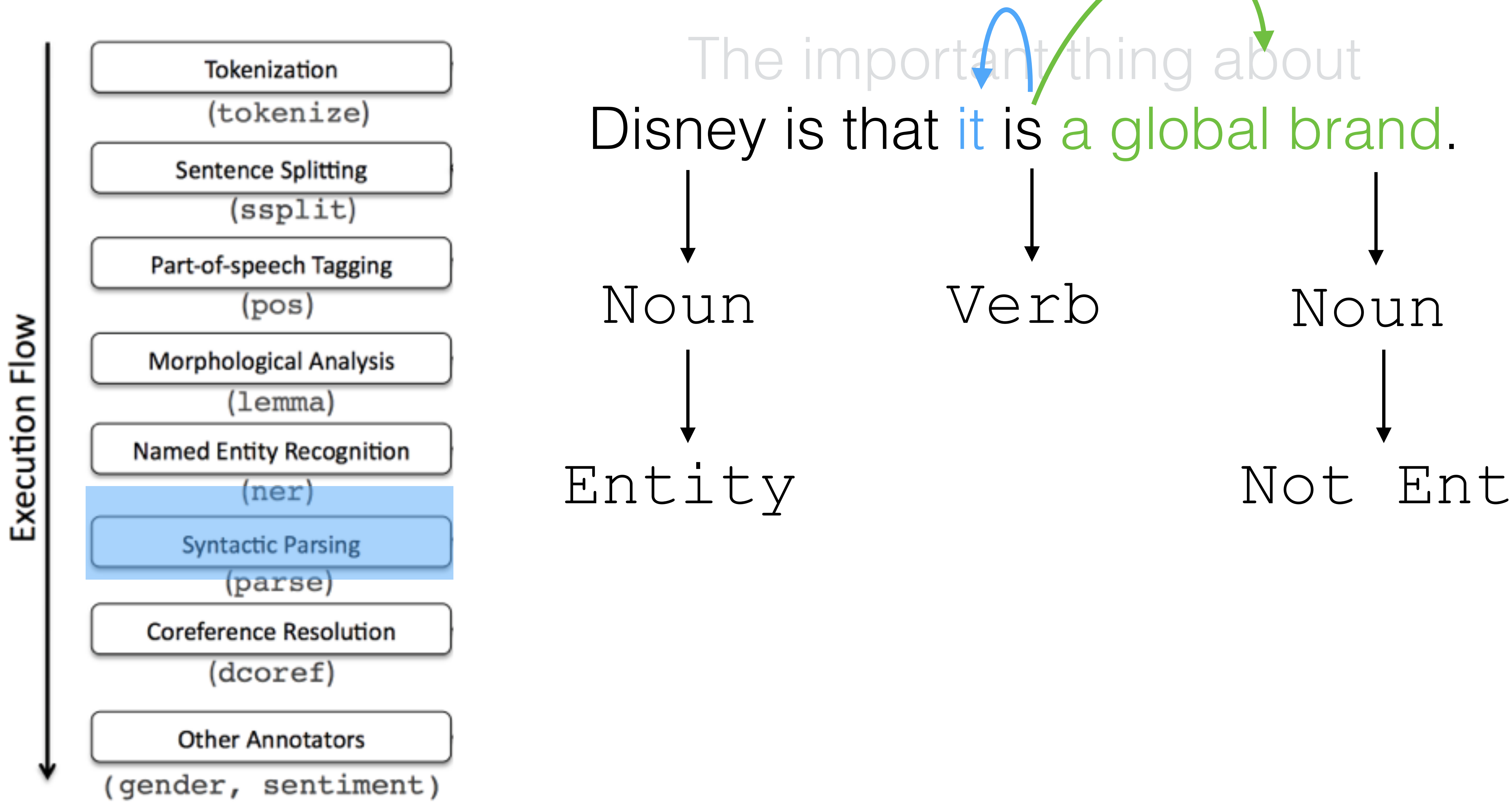
Is such-and-such feature encoded by the representation?





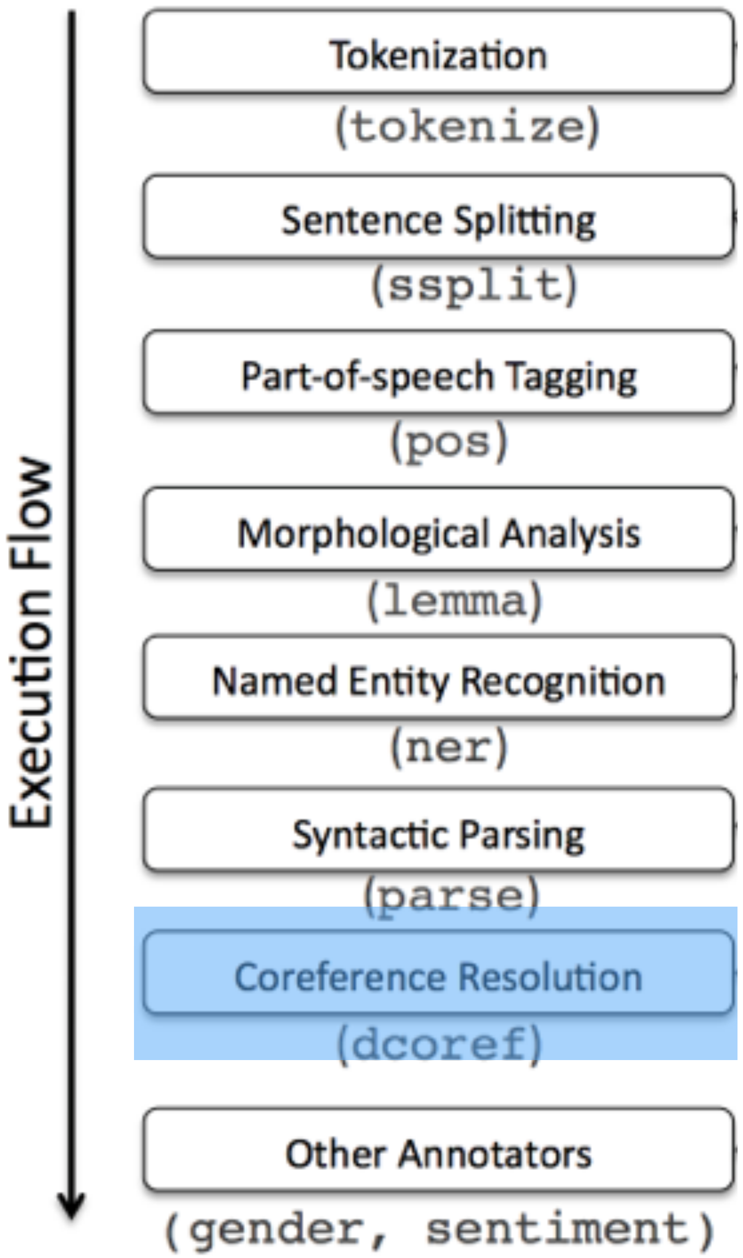
Probing Sentence Structure in Contextualized Word Representations.
Tenney et al. (ICLR 2019)

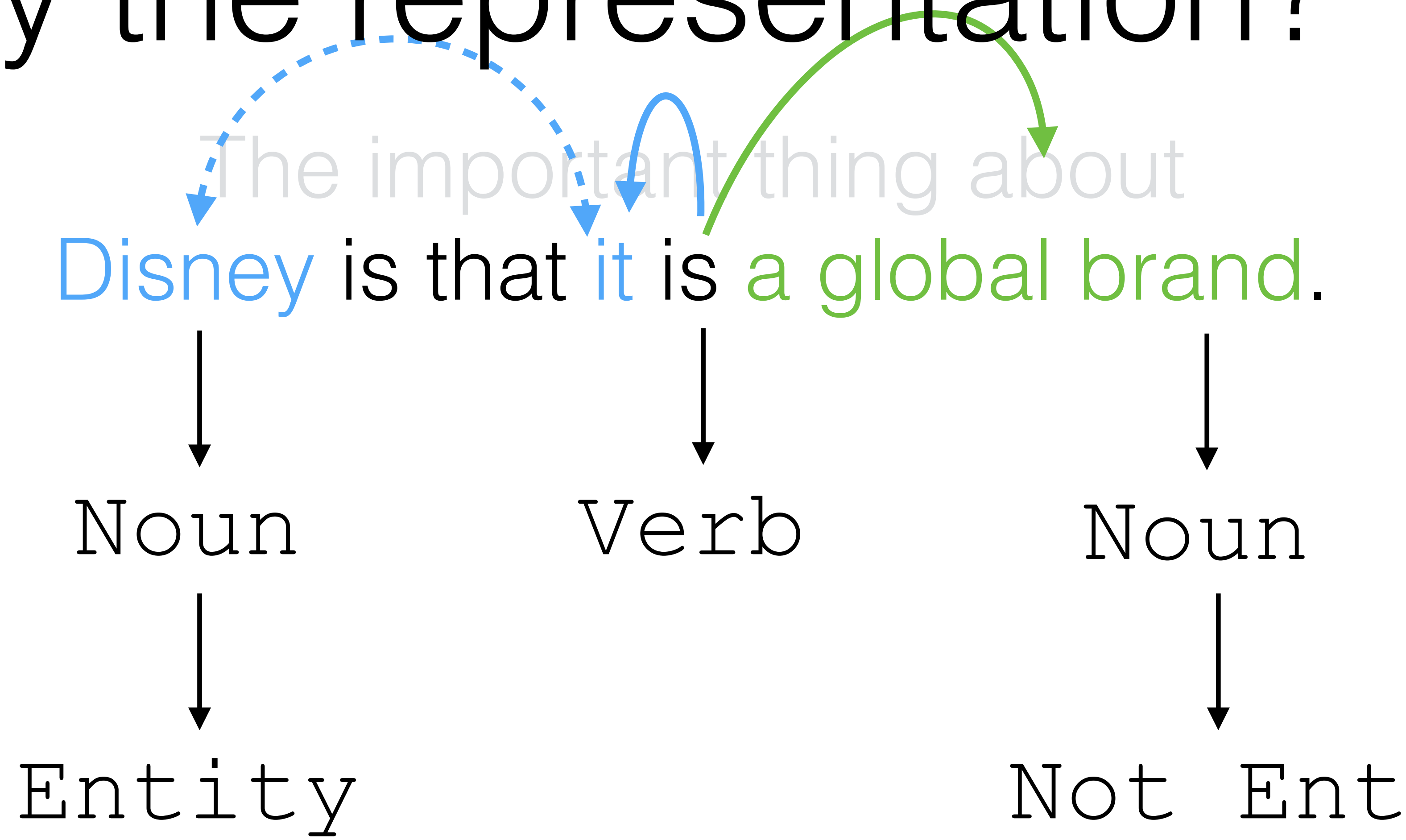
Is such-and-such feature encoded by the representation?



Probing Sentence Structure in Contextualized Word Representations.
Tenney et al. (ICLR 2019)

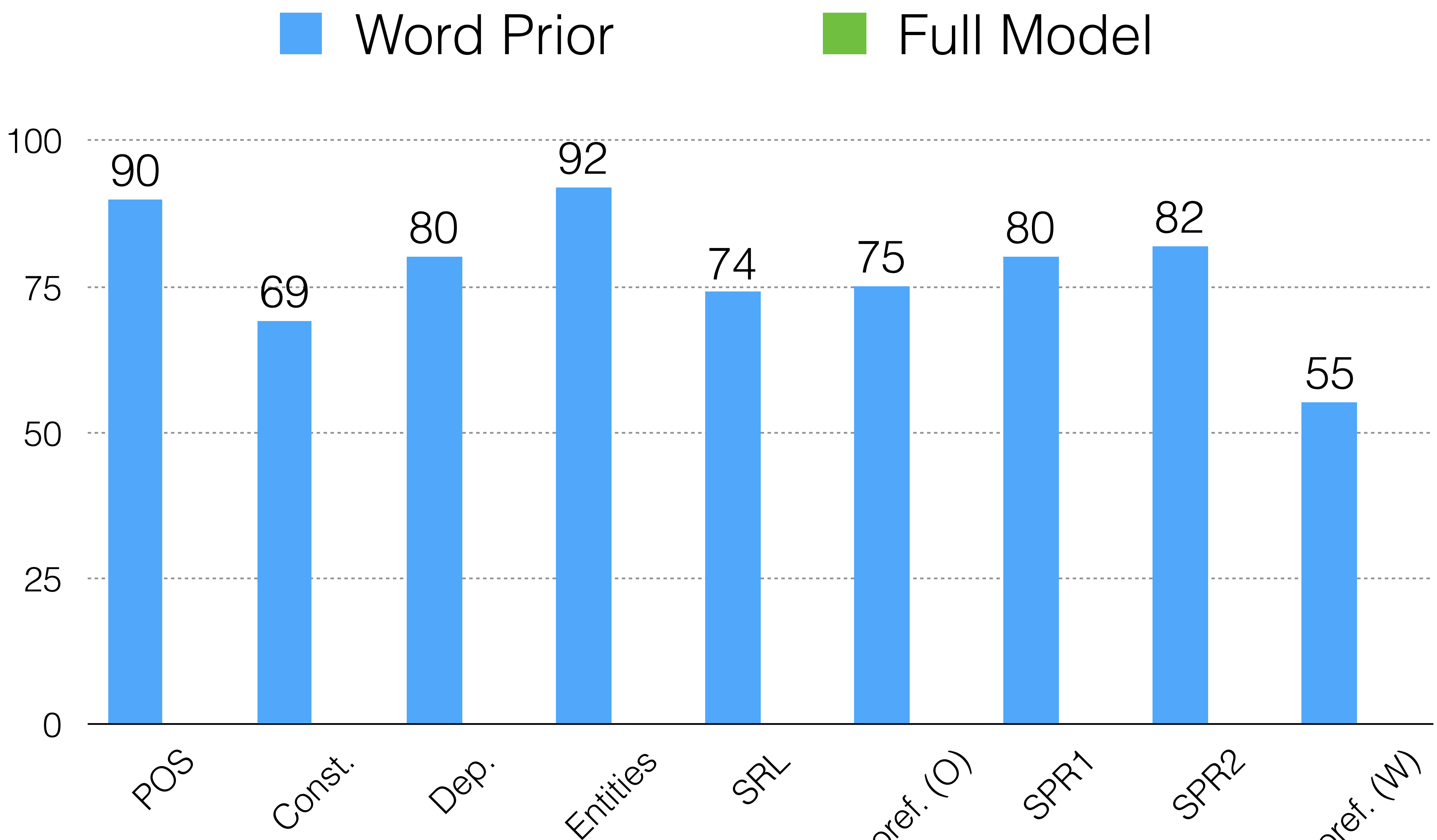
Is such-and-such feature encoded by the representation?





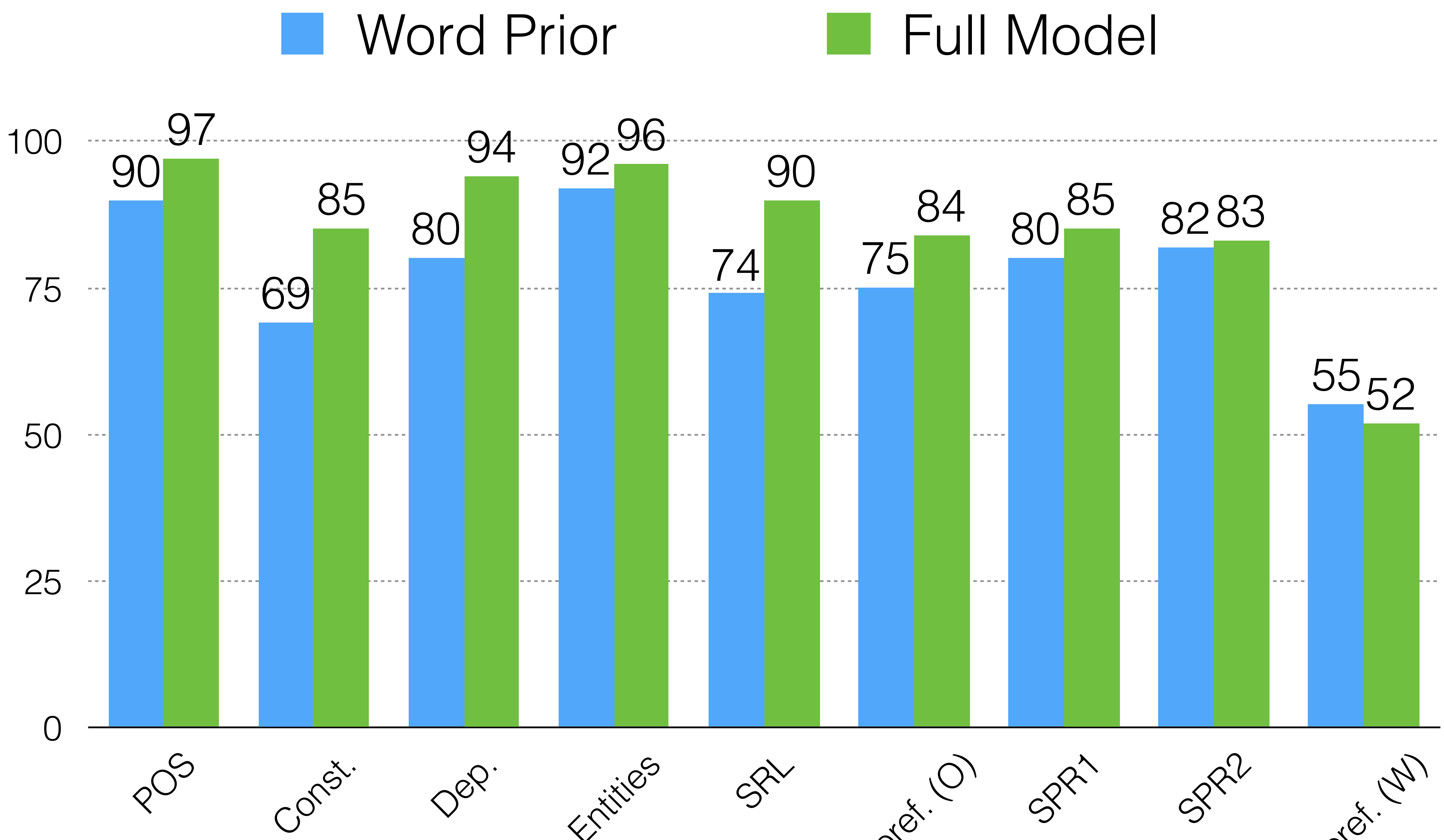
Probing Sentence Structure in Contextualized Word Representations.
Tenney et al. (ICLR 2019)

Is such-and-such feature encoded by the representation?



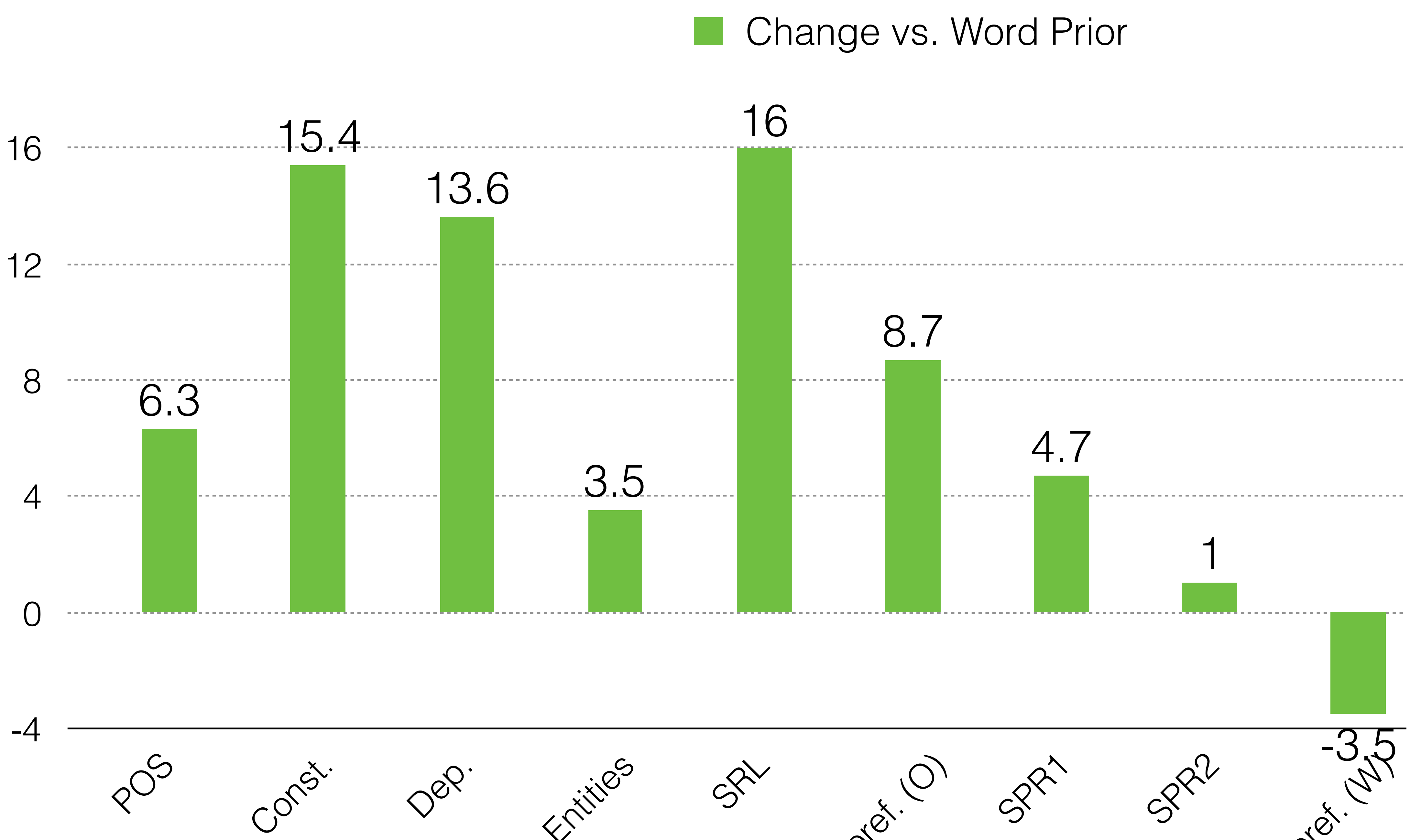
Probing Sentence Structure in Contextualized Word Representations.
Tenney et al. (ICLR 2019)

Is such-and-such feature encoded by the representation?



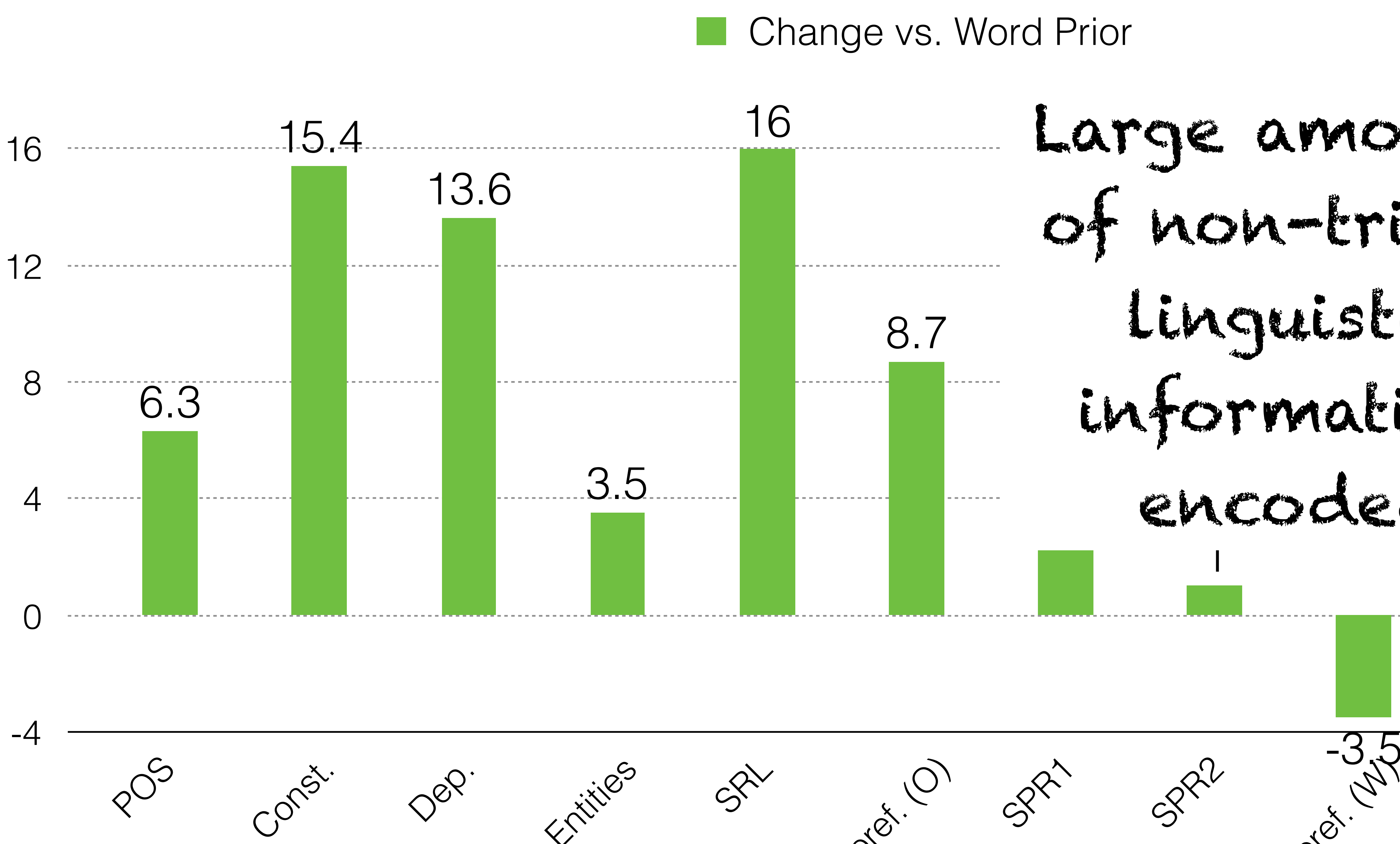
Probing Sentence Structure in Contextualized Word Representations.
Tenney et al. (ICLR 2019)

Is such-and-such feature encoded by the representation?



Probing Sentence Structure in Contextualized Word Representations.
Tenney et al. (ICLR 2019)

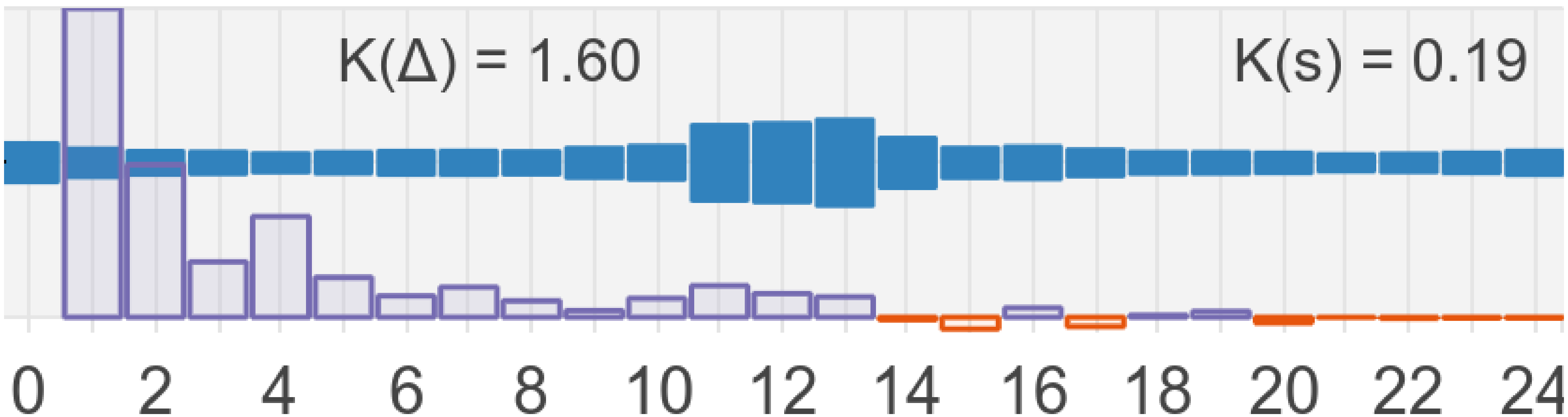
Is such-and-such feature encoded by the representation?



Large amounts of non-trivial linguistic information encoded

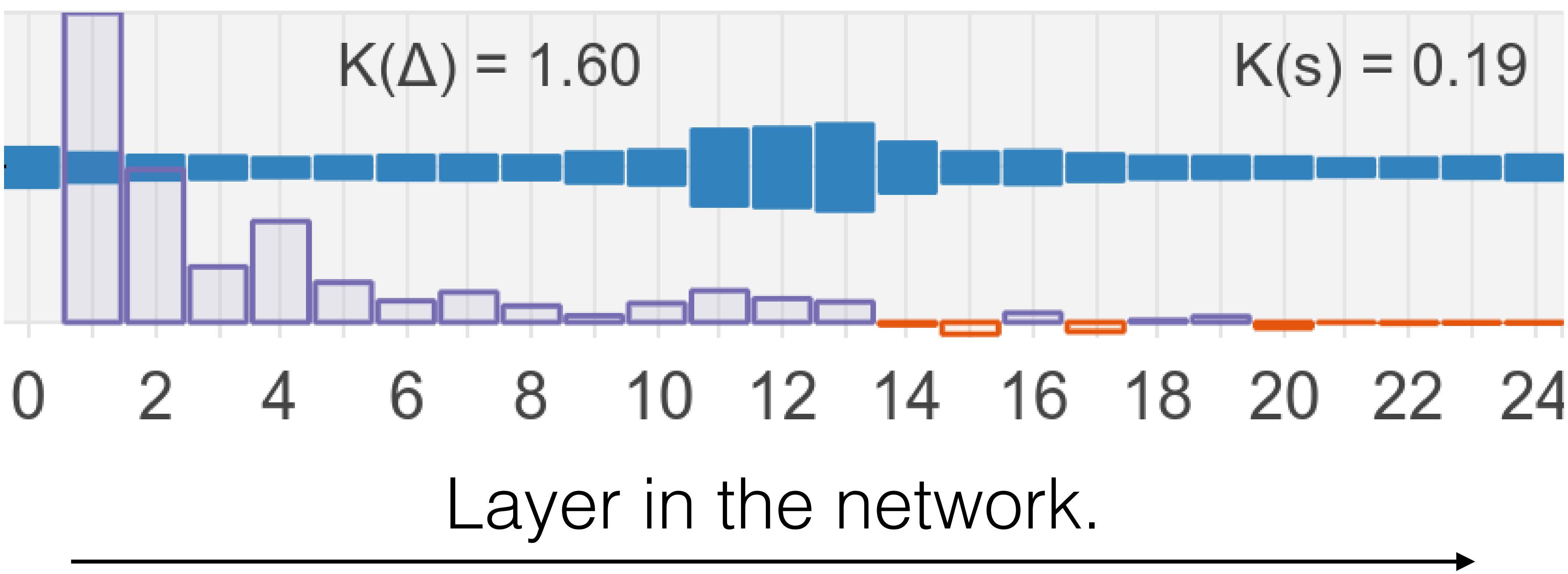
Probing Sentence Structure in Contextualized Word Representations.
Tenney et al. (ICLR 2019)

Is such-and-such feature encoded by the representation?



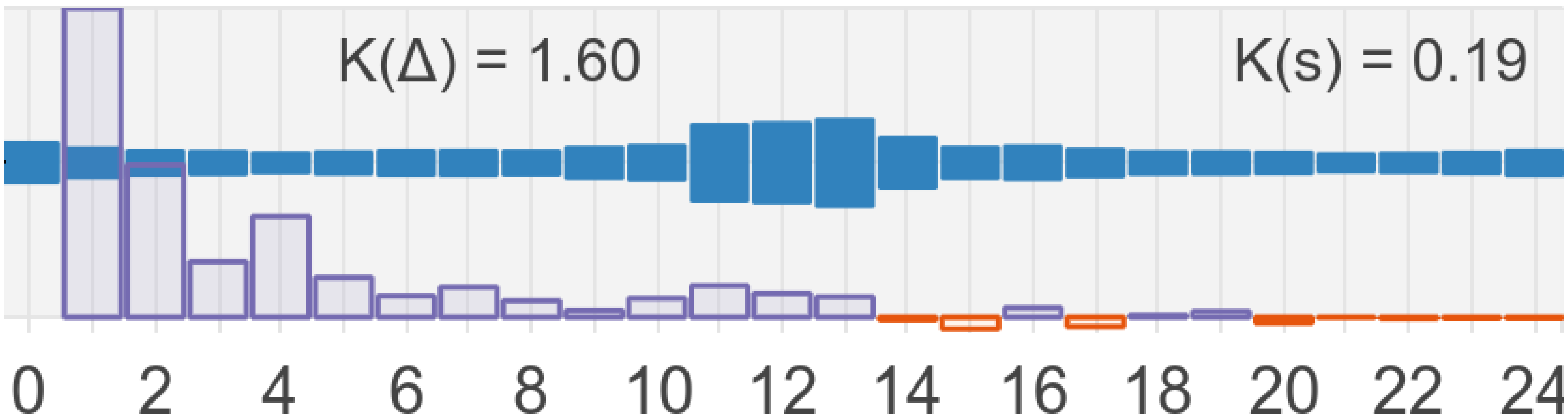
BERT Rediscovers the Classical NLP Pipeline
Tenney, Das, and Pavlick (ACL 2019)

Is such-and-such feature encoded by the representation?



BERT Rediscovers the Classical NLP Pipeline
Tenney, Das, and Pavlick (ACL 2019)

Is such-and-such feature encoded by the representation?



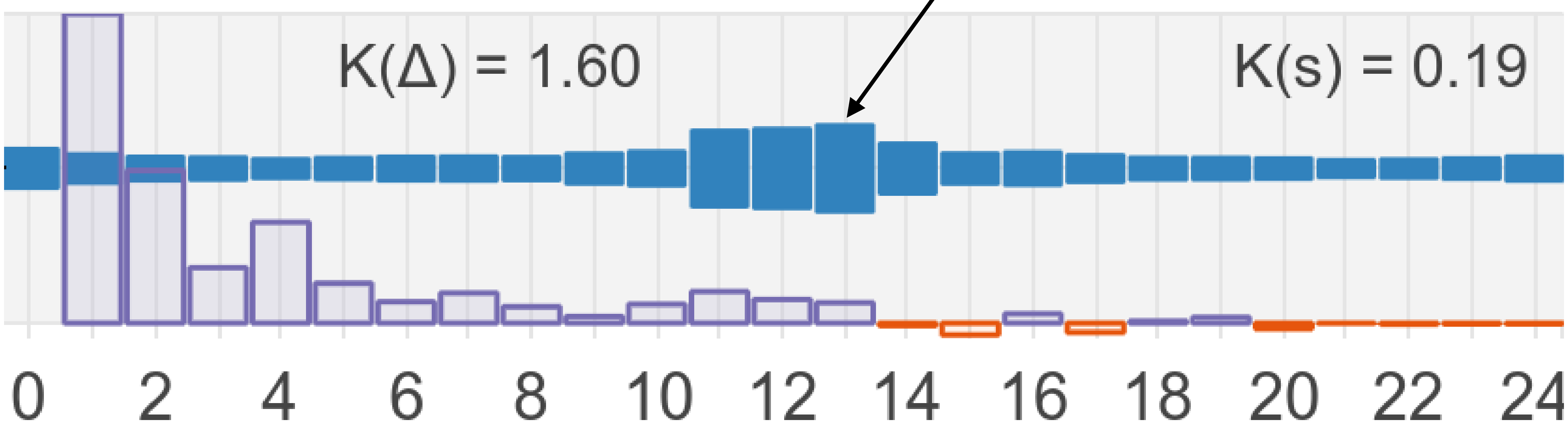
Layer in the network.

(Higher-level decisions can depend on lower-level ones.)

BERT Rediscovers the Classical NLP Pipeline
Tenney, Das, and Pavlick (ACL 2019)

Is such-and-such feature encoded by the representation?

Importance of layer in decision

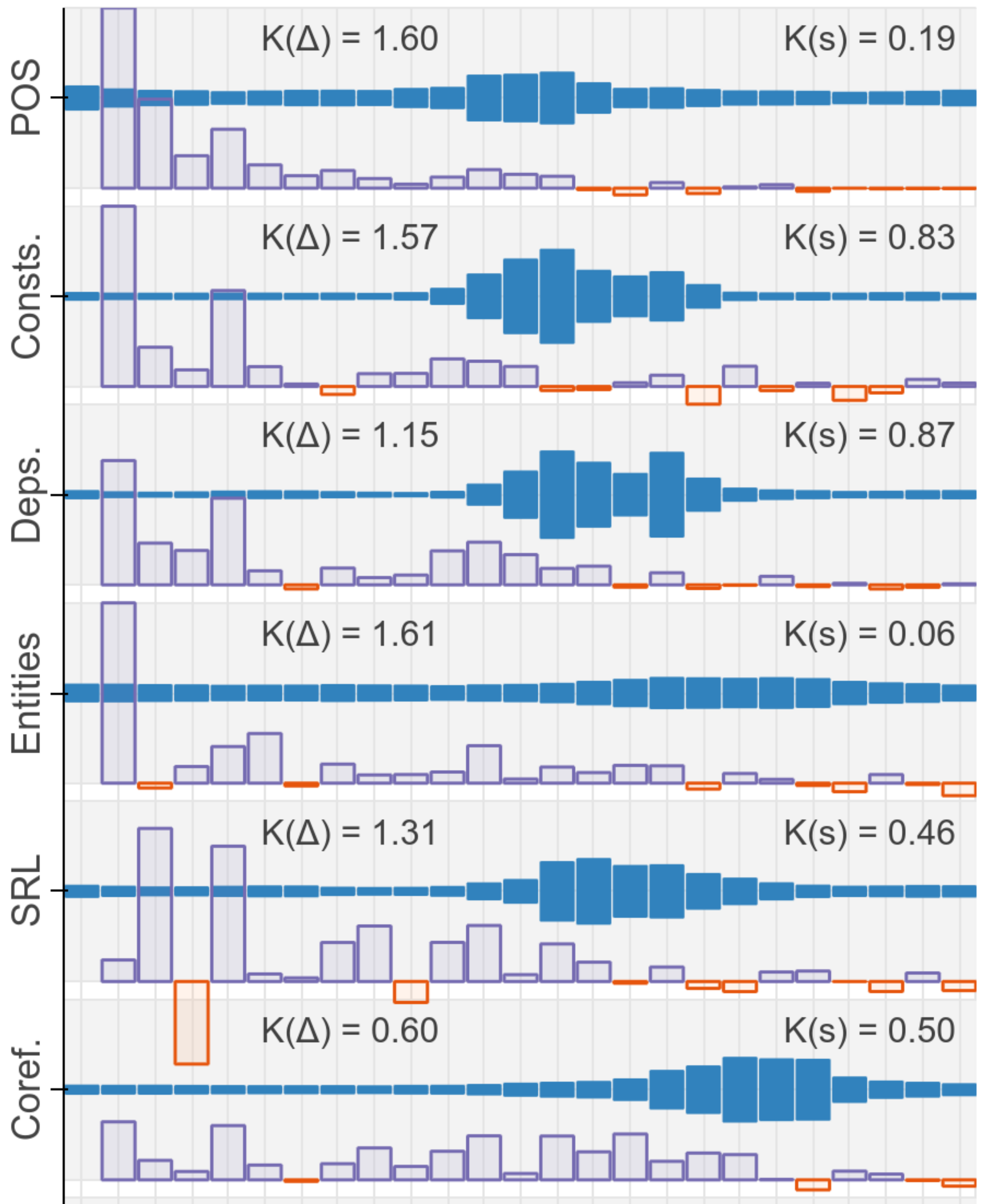


Layer in the network.

(Higher-level decisions can depend on lower-level ones.)

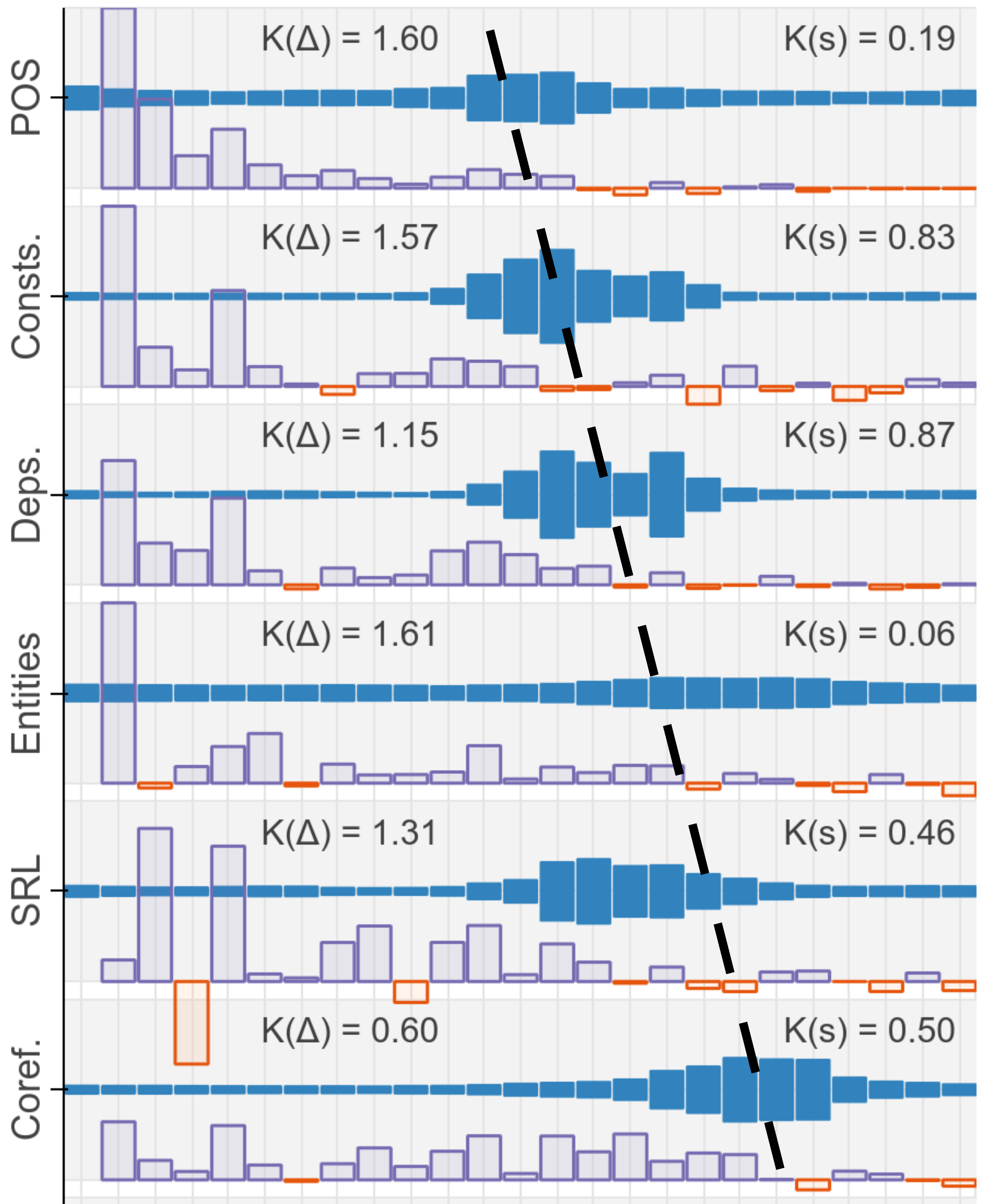
BERT Rediscovers the Classical NLP Pipeline
Tenney, Das, and Pavlick (ACL 2019)

Is such-and-such feature encoded by the representation?



BERT Rediscovers the Classical NLP Pipeline
Tenney, Das, and Pavlick (ACL 2019)

Is such-and-such feature encoded by the representation?



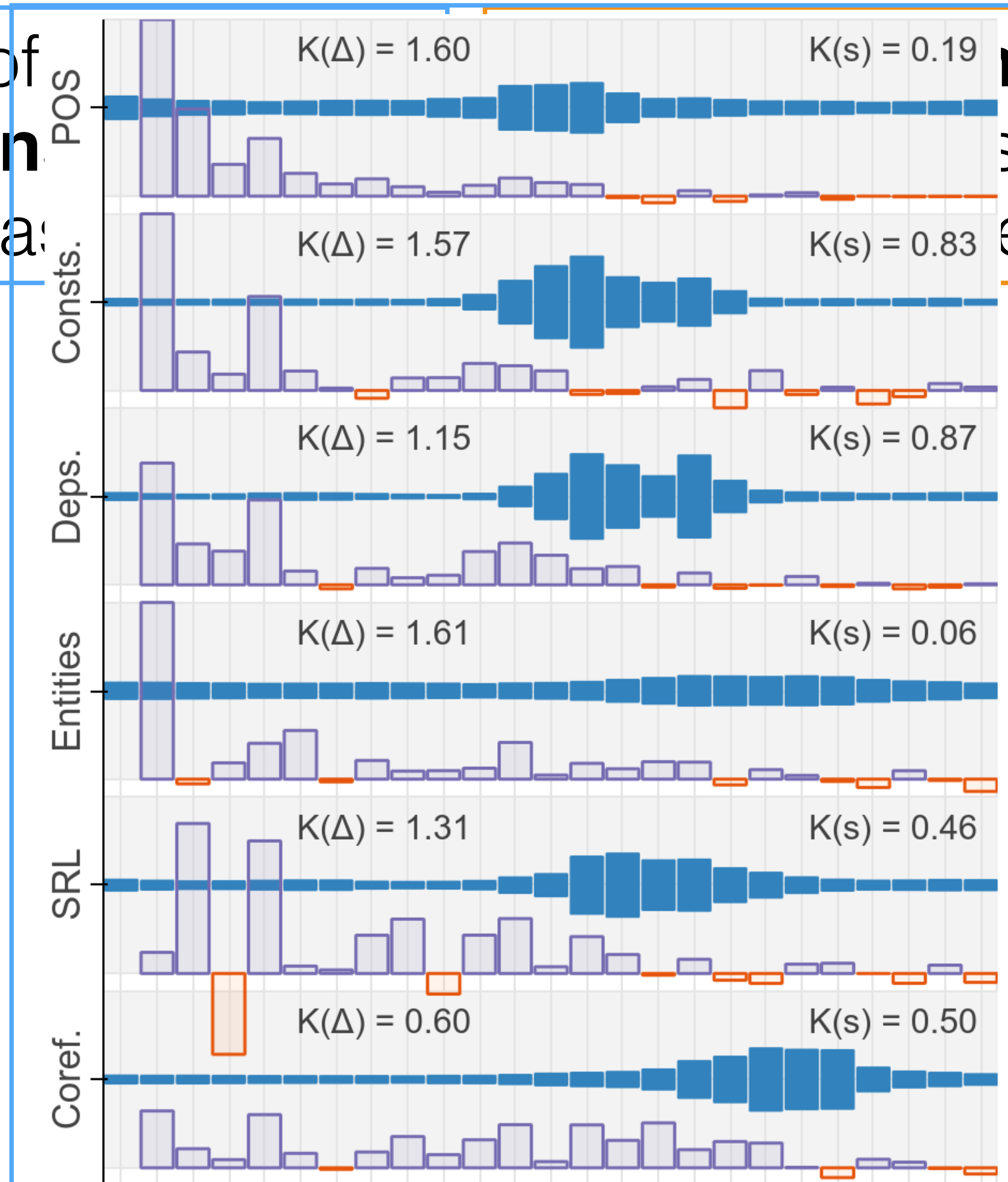
Roughly: Higher-level information gets encoded later in the network.

BERT Rediscovers the Classical NLP Pipeline
Tenney, Das, and Pavlick (ACL 2019)

Past ~2 years:
What do deep LMs know about language?

What types of
representation
(“Probing Cla

ave like they
se features?
e Tasks”)

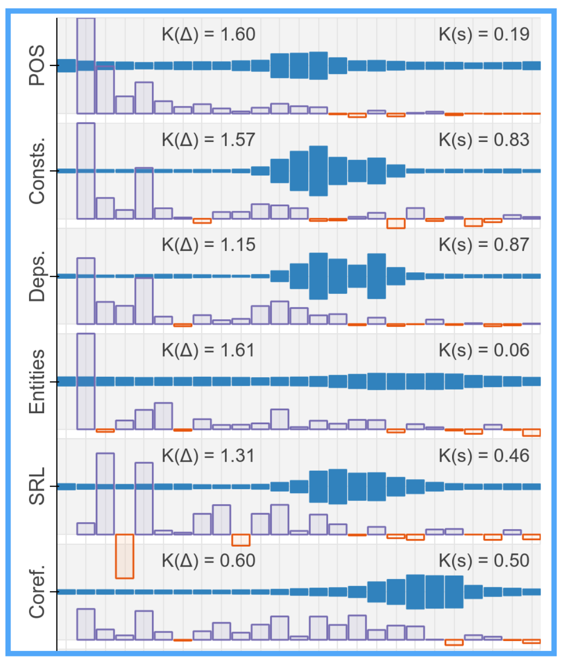


Tenney et al (ACL 2019)

Past ~2 years:
What do deep LMs know about language?

What types of features **representations** encode?
("Probing Classifiers")

Do models **behave** like they are using these features?
("Challenge Tasks")

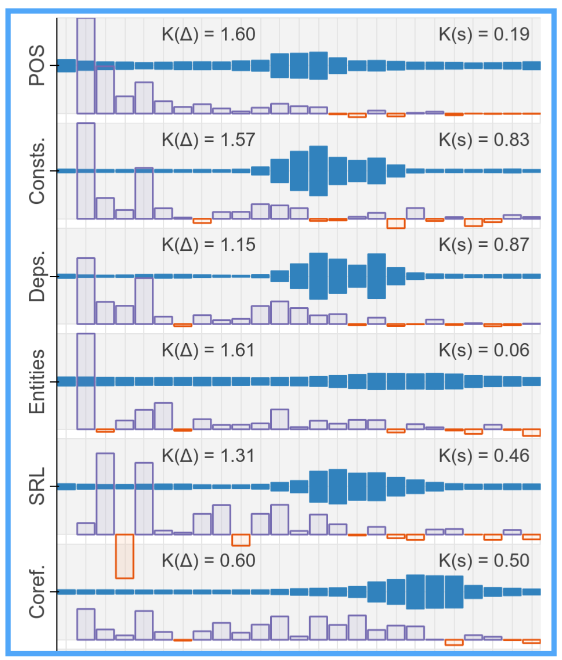


Tenney et al (ACL 2019)

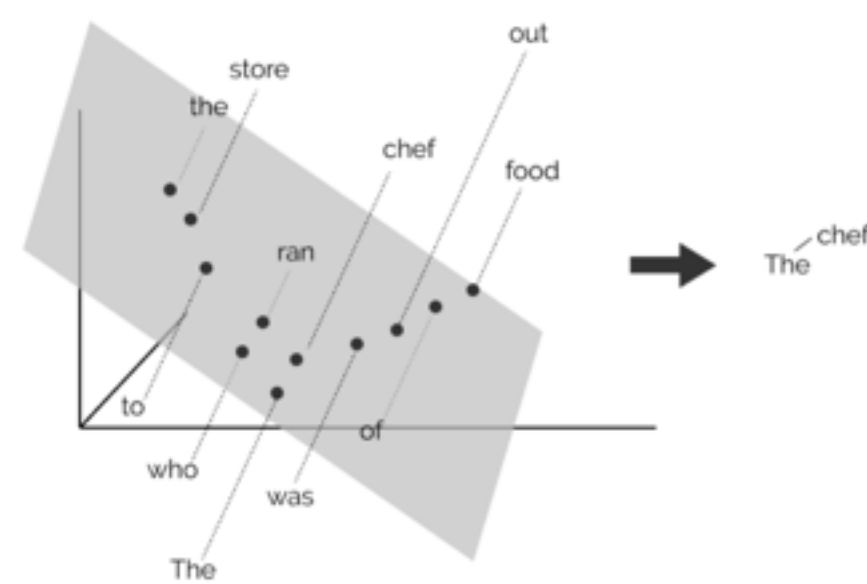
Past ~2 years:
What do deep LMs know about language?

What types of features **representations** encode?
("Probing Classifiers")

Do models **behave** like they are using these features?
("Challenge Tasks")



Tenney et al (ACL 2019)

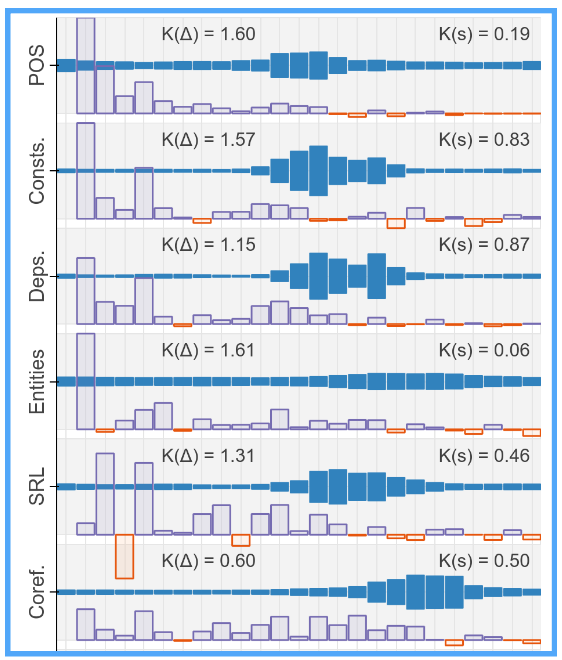


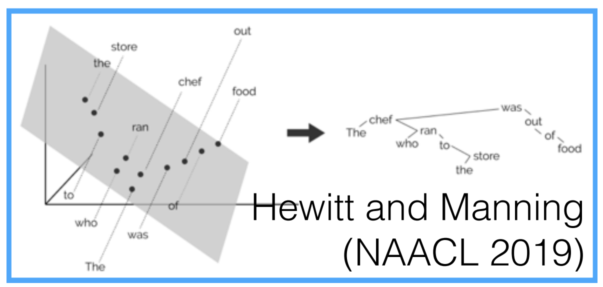
Hewitt and Manning
(NAACL 2019)

Past ~2 years:
What do deep LMs know about language?

What types of features **representations** encode?
("Probing Classifiers")

Do models **behave** like they are using these features?
("Challenge Tasks")





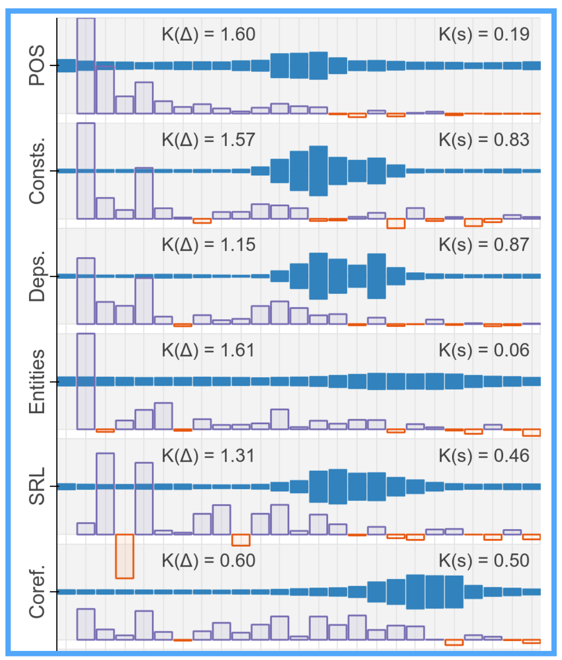
Hewitt and Manning (NAACL 2019)

Tenney et al (ACL 2019)

Past ~2 years:
What do deep LMs know about language?

What types of features **representations** encode?
(“Probing Classifiers”)

Do models **behave** like they are using these features?
(“Challenge Tasks”)



Tenney et al (ACL 2019)

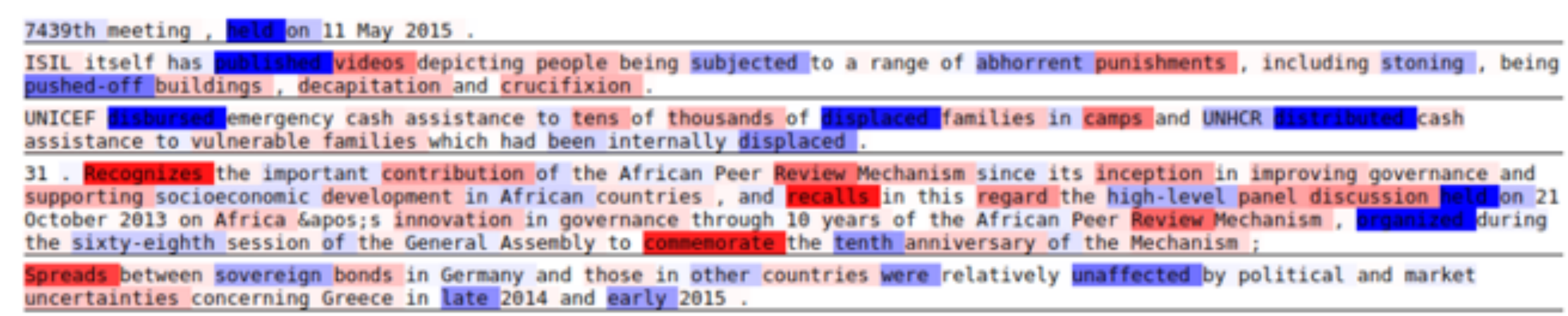


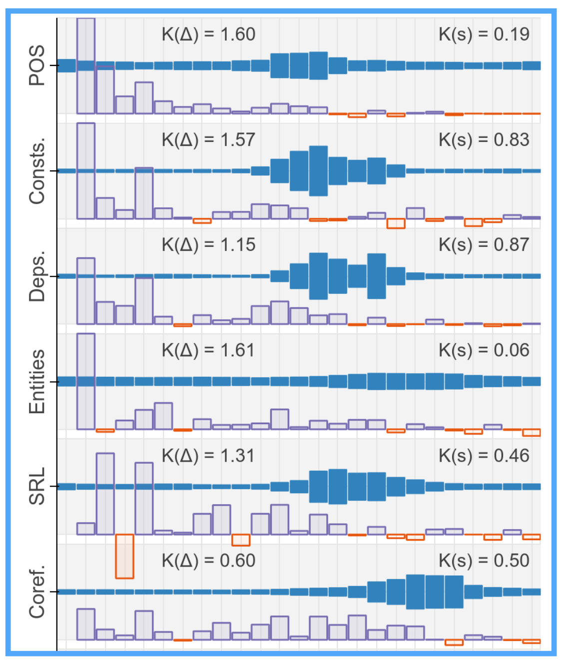
Figure 5: Visualization of a neuron from an English-Arabic model that activates on verb tense: negative/positive for past/present. Examples shown are the first 5 sentences in the test set.

Bau et al. (ICLR 2019)

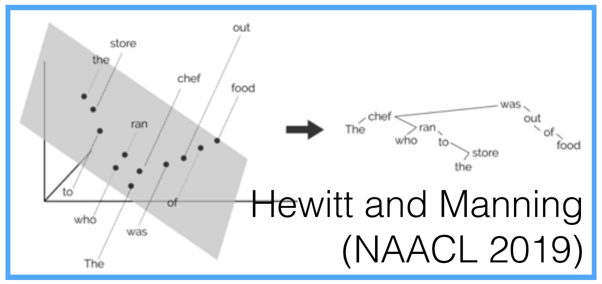
Past ~2 years:
What do deep LMs know about language?

What types of features **representations** encode?
(“Probing Classifiers”)

Do models **behave** like they are using these features?
(“Challenge Tasks”)



Tenney et al (ACL 2019)



Hewitt and Manning
(NAACL 2019)

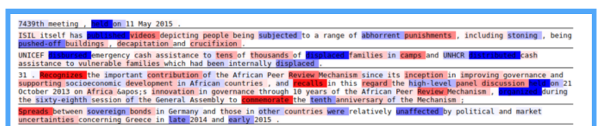


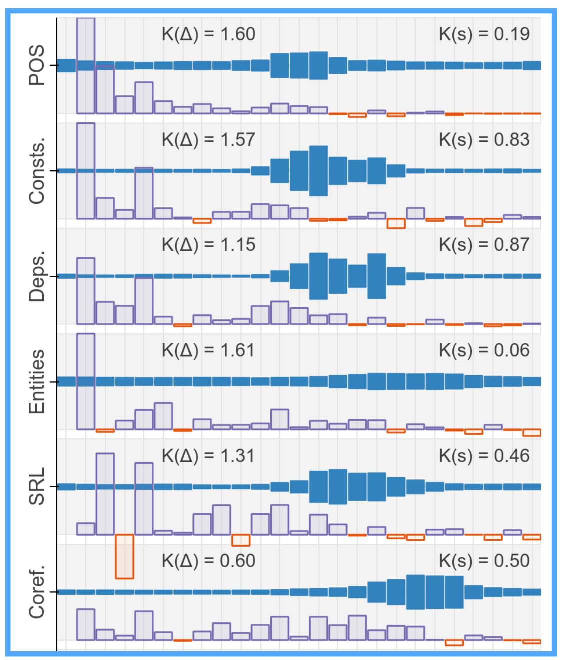
Figure 5: Visualization of a neuron from an English-Arabic model that activates on verb tense: negative/positive for past/present. Examples shown are the first 5 sentences in the test set.

Bau et al. (ICLR 2019)

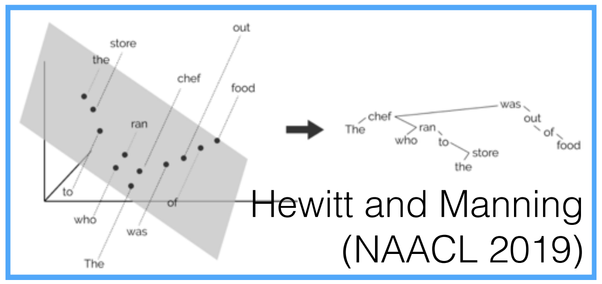
Past ~2 years:
What do deep LMs know about language?

What types of features **representations** encode?
("Probing Classifiers")

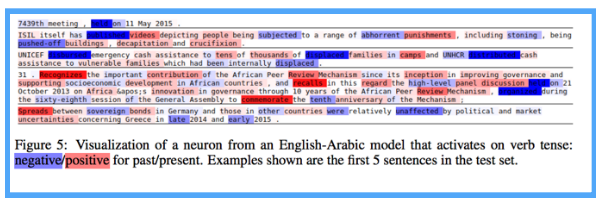
Do models **behave** like they are using these features?
("Challenge Tasks")



Tenney et al (ACL 2019)



Hewitt and Manning (NAACL 2019)



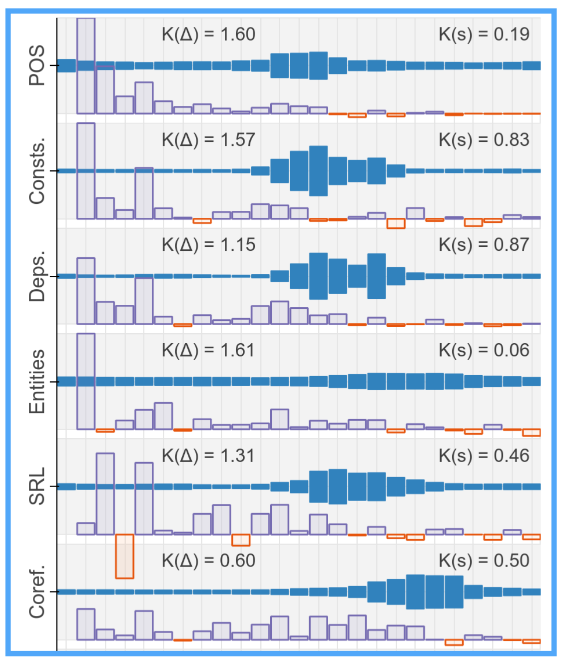
Bau et al. (ICLR 2019)

Wealth of evidence that linguistic information is "there"

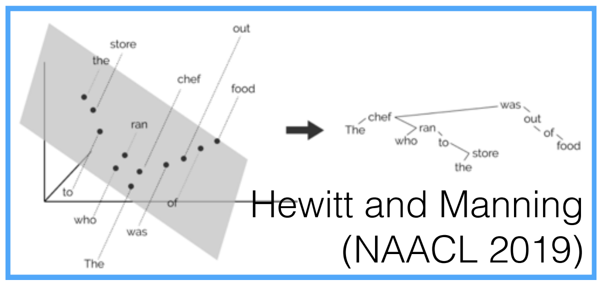
Past ~2 years:
What do deep LMs know about language?

What types of features **representations** encode?
("Probing Classifiers")

Do models **behave** like they are using these features?
("Challenge Tasks")



Tenney et al (ACL 2019)



Hewitt and Manning (NAACL 2019)

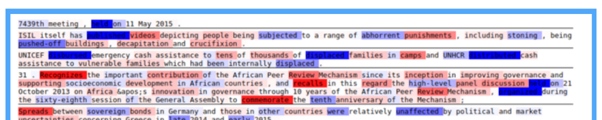


Figure 5: Visualization of a neuron from an English-Arabic model that activates on verb tense: negative/positive for past/present. Examples shown are the first 5 sentences in the test set.

Bau et al. (ICLR 2019)

Wealth of evidence that linguistic information is "there"

Is such-and-such feature used by the model?

Is such-and-such feature used by the model?

There are apples and bananas on the table.

There are apples on the table.

Is such-and-such feature used by the model?

Premise

There are apples and bananas on the table.

Hypothesis

There are apples on the table.

Is such-and-such feature used by the model?

Premise

There are apples and bananas on the table.

Hypothesis

There are apples on the table.

Is such-and-such feature used by the model?

Premise

There are apples or bananas on the table.

Hypothesis

There are apples on the table.

Is such-and-such feature used by the model?

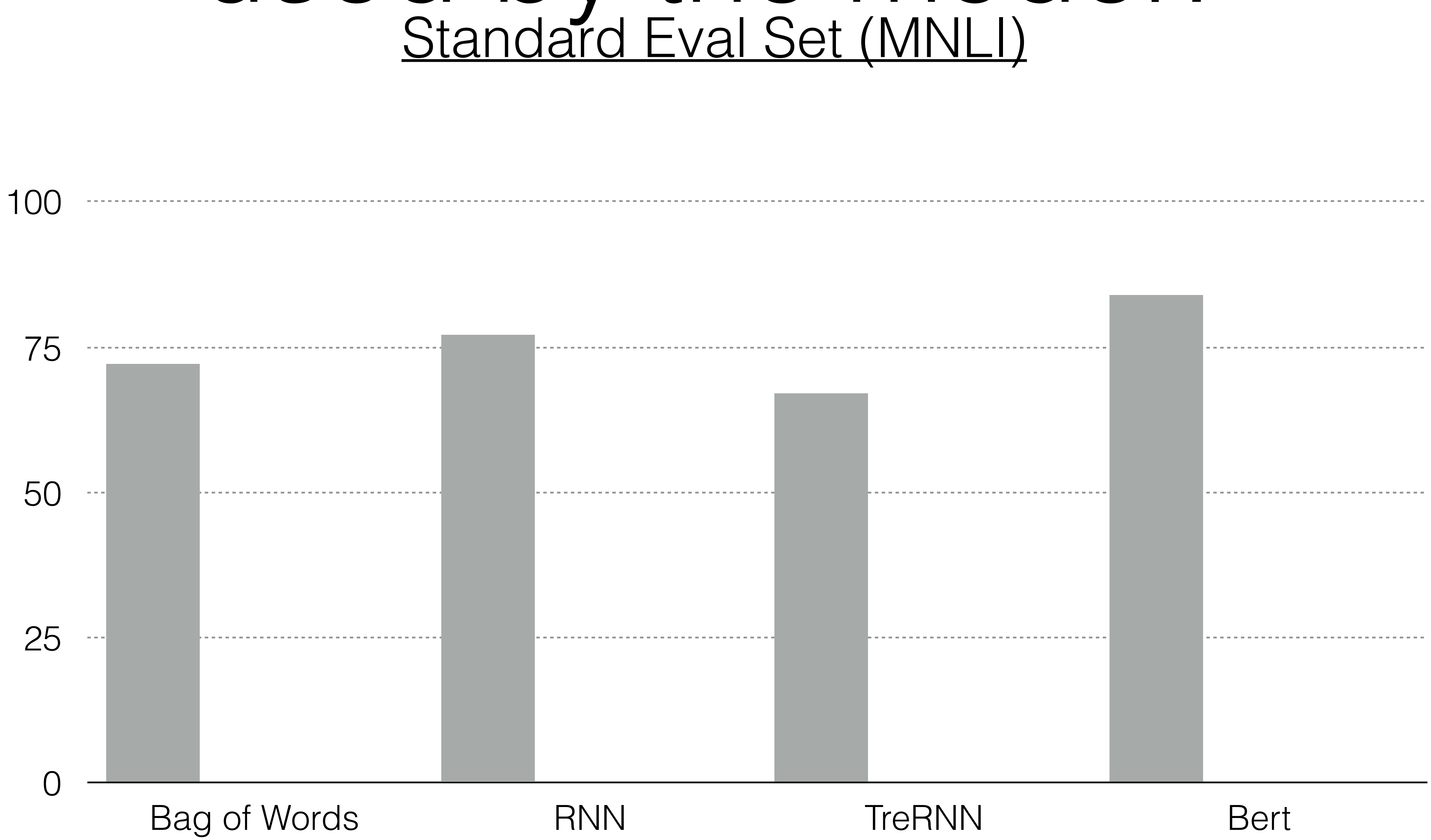
Lexical Overlap Heuristic

The banker near the judge saw the actor.
The banker saw the actor.

The judge by the actor stopped the banker.
The banker stopped the judge.

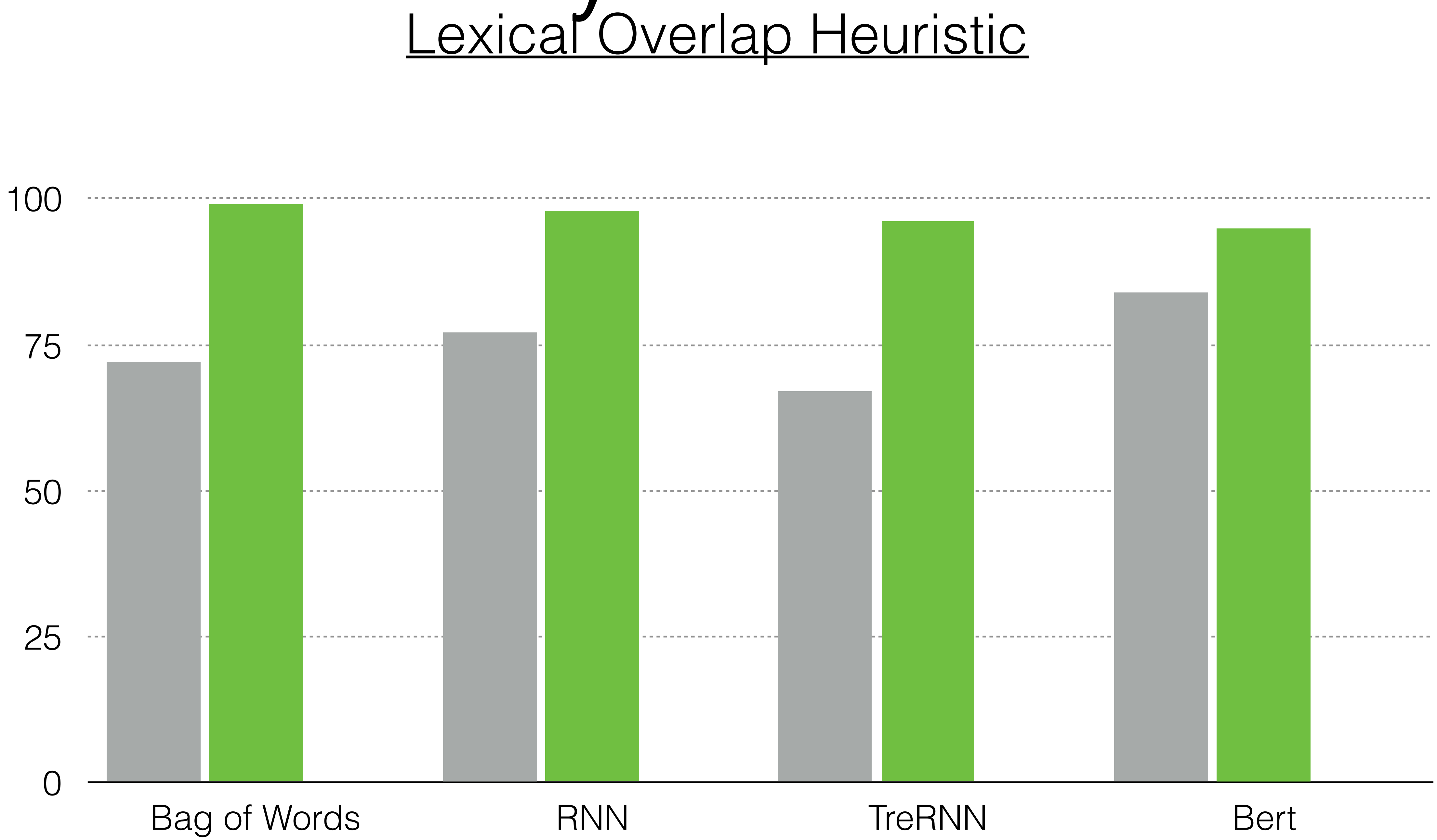
Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference.
McCoy, Pavlick, and Linzen (2019)

Is such-and-such feature used by the model?



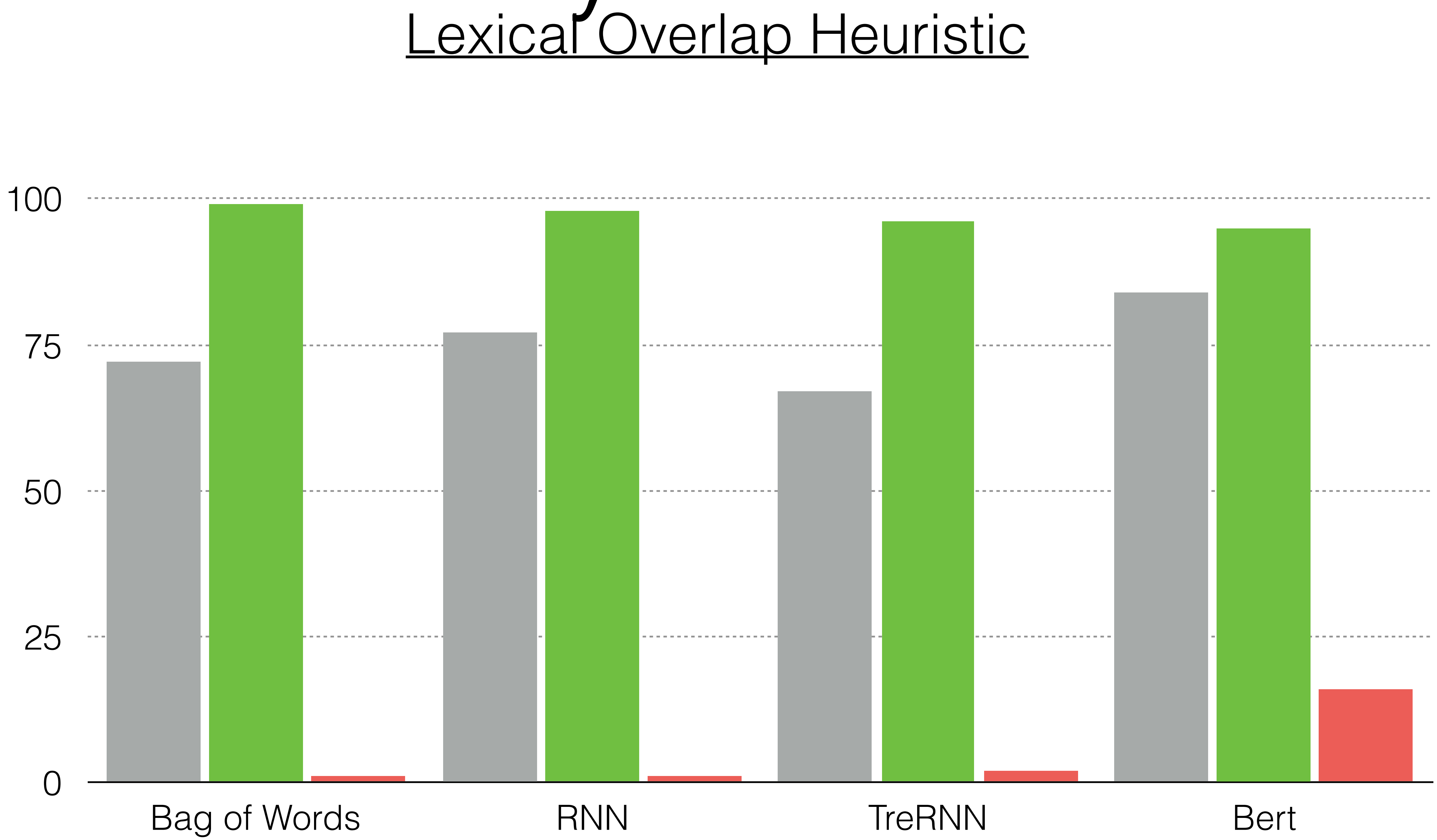
Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference.
McCoy, Pavlick, and Linzen (2019)

Is such-and-such feature used by the model?



Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference.
McCoy, Pavlick, and Linzen (2019)

Is such-and-such feature used by the model?

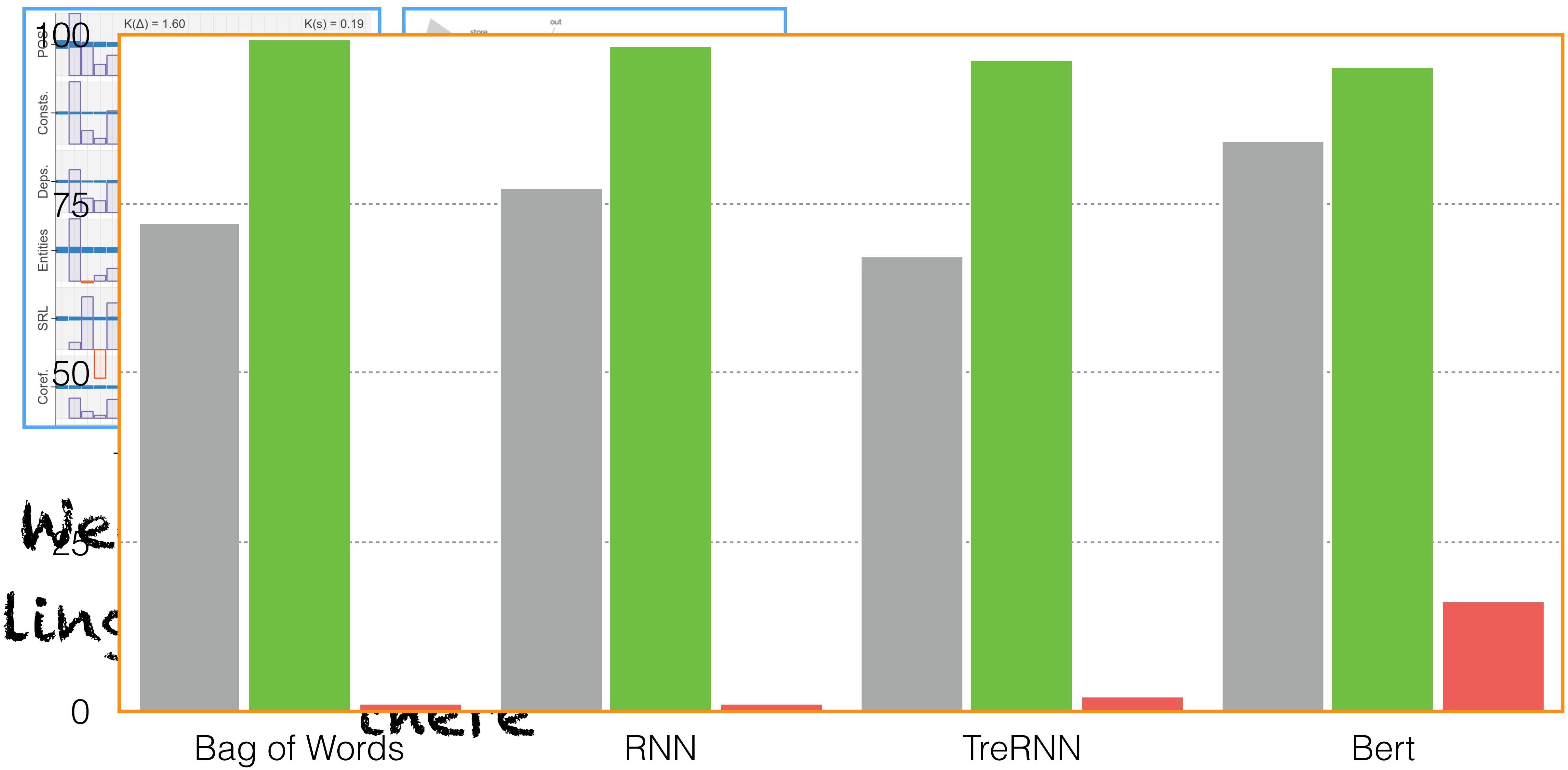


Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference.
McCoy, Pavlick, and Linzen (2019)

Past ~2 years:
What do deep LMs know about language?

What types of features **representations** encode?
("Probing Classifiers")

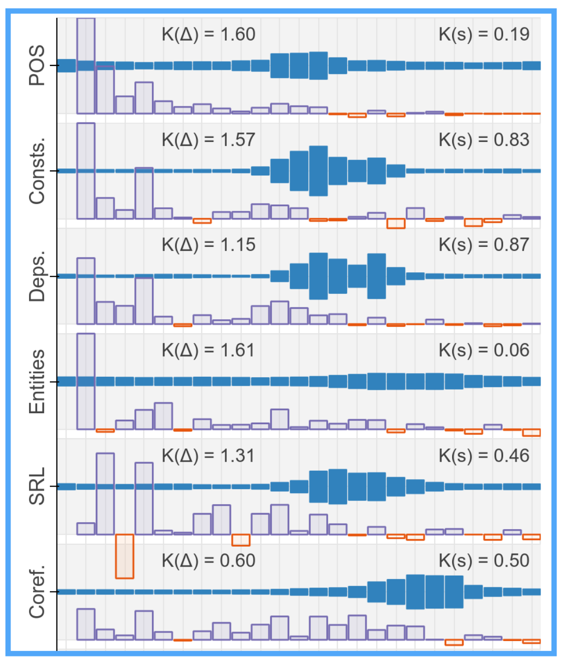
Do models **behave** like they are using these features?
("Challenge Tasks")



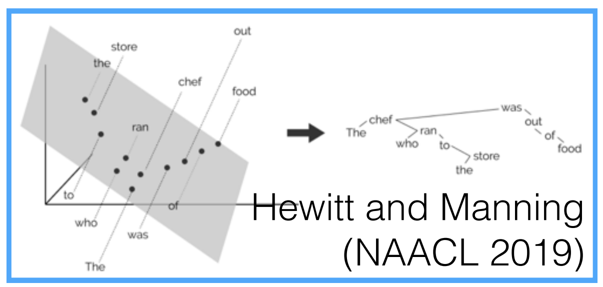
Past ~2 years:
What do deep LMs know about language?

What types of features **representations** encode?
("Probing Classifiers")

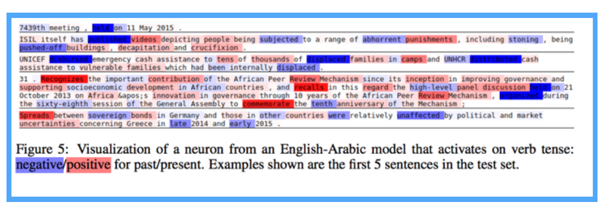
Do models **behave** like they are using these features?
("Challenge Tasks")



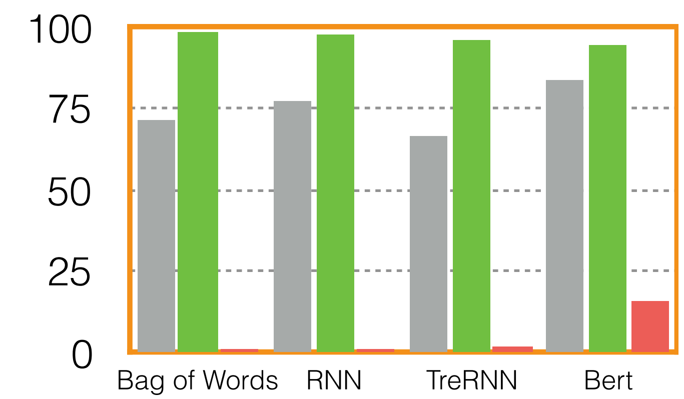
Tenney et al (ACL 2019)



Hewitt and Manning (NAACL 2019)



Bau et al. (ICLR 2019)



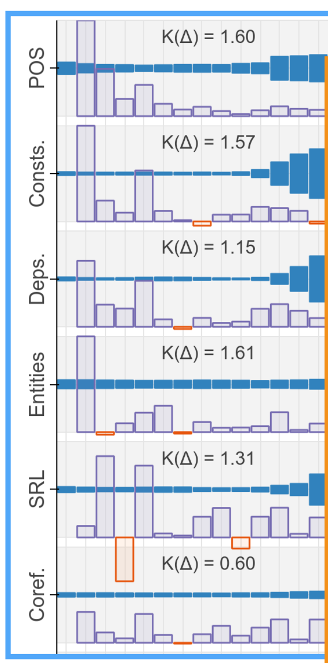
McCoy et al. (2019)

Wealth of evidence that linguistic information is "there"

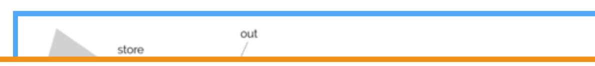
Past ~2 years:
What do deep LMs know about language?

What types of features **representations** encode?
("Probing Classifiers")

Do models **behave** like they are using these features?
("Challenge Tasks")



Tenney et a

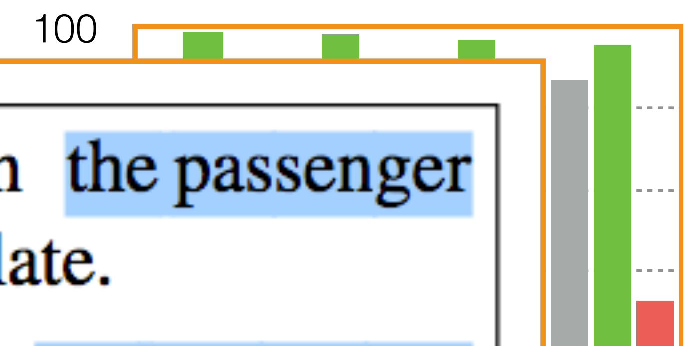


(1a) **The paramedic** performed CPR on the passenger even though she/he/they knew it was too late.

(2a) The paramedic performed CPR on **the passenger** even though she/he/they was/were already dead.

(1b) **The paramedic** performed CPR on someone even though she/he/they knew it was too late.

(2b) The paramedic performed CPR on **someone** even though she/he/they was/were already dead.



Bert
(2019)

Wealth

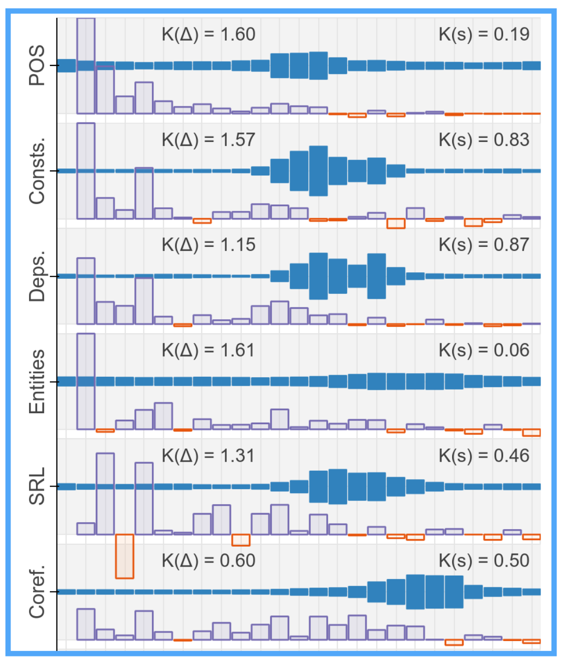
Linguis

"there"

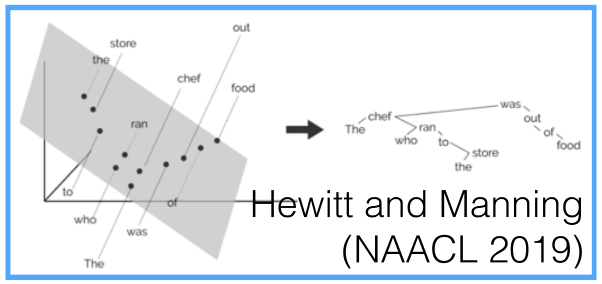
Past ~2 years:
What do deep LMs know about language?

What types of features **representations** encode?
(“Probing Classifiers”)

Do models **behave** like they are using these features?
(“Challenge Tasks”)



Tenney et al (ACL 2019)



Hewitt and Manning (NAACL 2019)

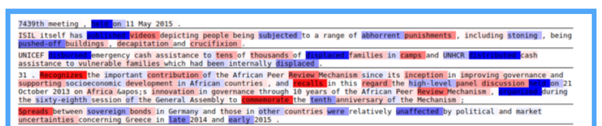
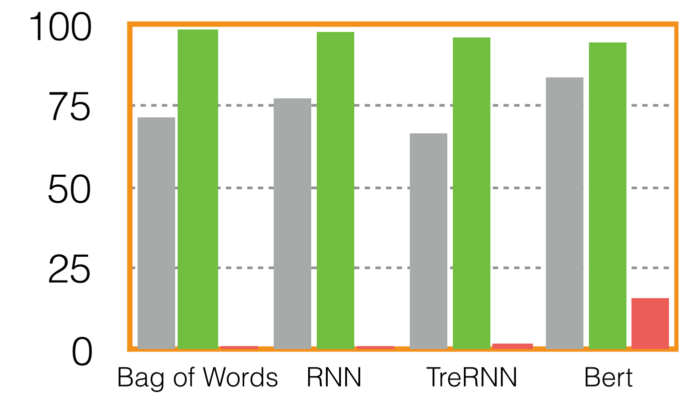


Figure 5: Visualization of a neuron from an English-Arabic model that activates on verb tense: negative/positive for past/present. Examples shown are the first 5 sentences in the test set.

Bau et al. (ICLR 2019)



McCoy et al. (2019)

(1a) **The paramedic** performed CPR on the passenger even though she/he/they knew it was too late.

(2a) The paramedic performed CPR on **the passenger** even though she/he/they was/were already dead.

(1b) **The paramedic** performed CPR on someone even though she/he/they knew it was too late.

(2b) The paramedic performed CPR on **someone** even though she/he/they was/were already dead.

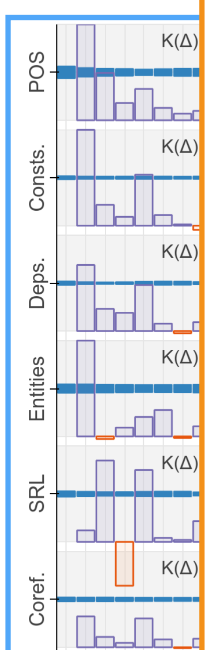
Rudinger et al. (2018)

Wealth of evidence that linguistic information is “there”

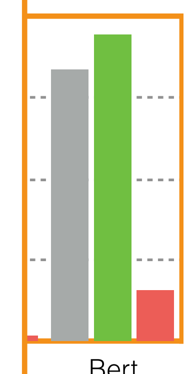
Past ~2 years:

W … e?

W … rep … (… ey … ?



Ten



l. (2019)

passenger
passenger
someone
someone

. (2018)

Article: Super Bowl 50

Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"

Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"

Original Prediction: John Elway

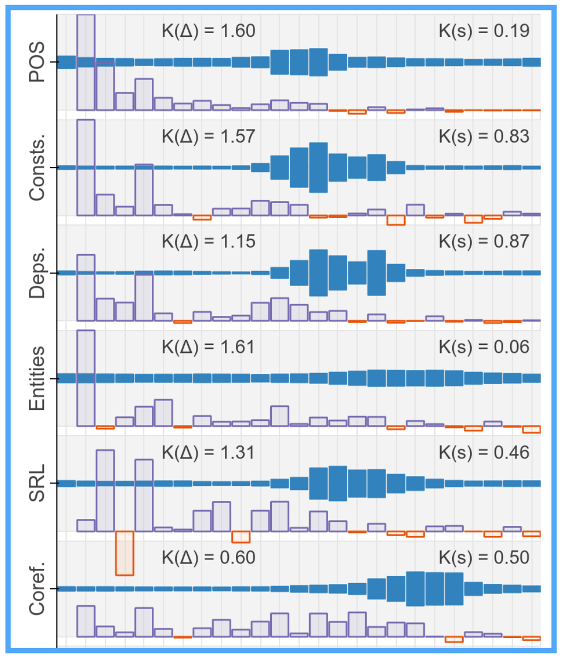
Prediction under adversary: Jeff Dean

Wea … Linguistic information is "there"

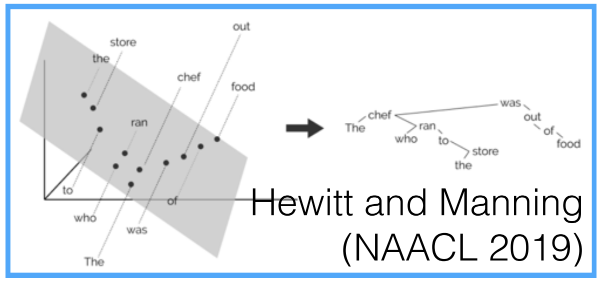
Past ~2 years:
What do deep LMs know about language?

What types of features **representations** encode?
("Probing Classifiers")

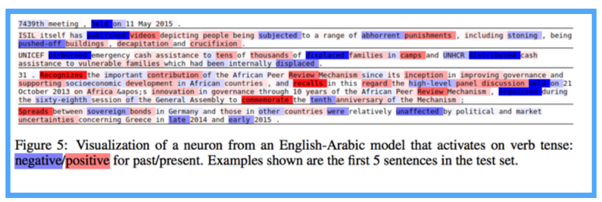
Do models **behave** like they are using these features?
("Challenge Tasks")



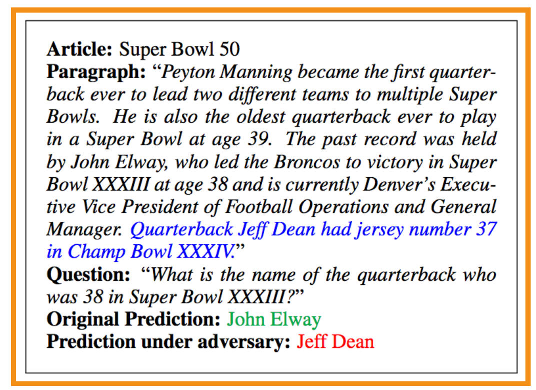
Tenney et al (ACL 2019)



Hewitt and Manning (NAACL 2019)

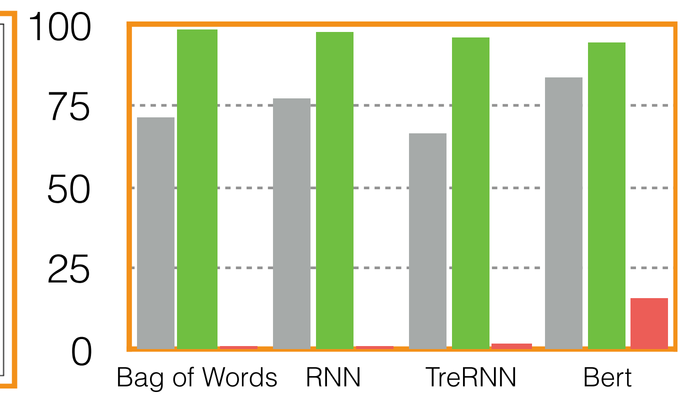


Bau et al. (ICLR 2019)

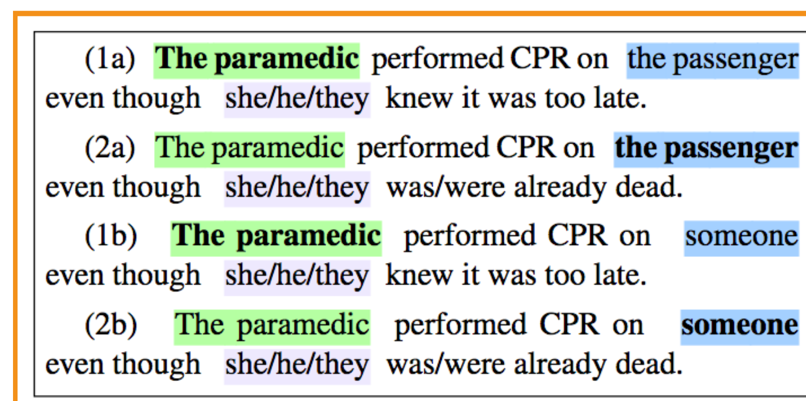


Article: Super Bowl 50
Paragraph: *"Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV."*
Question: *"What is the name of the quarterback who was 38 in Super Bowl XXXIII?"*
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Jia and Liang (2017)



McCoy et al. (2019)



(1a) **The paramedic** performed CPR on the passenger even though she/he/they knew it was too late.

(2a) The paramedic performed CPR on **the passenger** even though she/he/they was/were already dead.

(1b) **The paramedic** performed CPR on someone even though she/he/they knew it was too late.

(2b) The paramedic performed CPR on **someone** even though she/he/they was/were already dead.

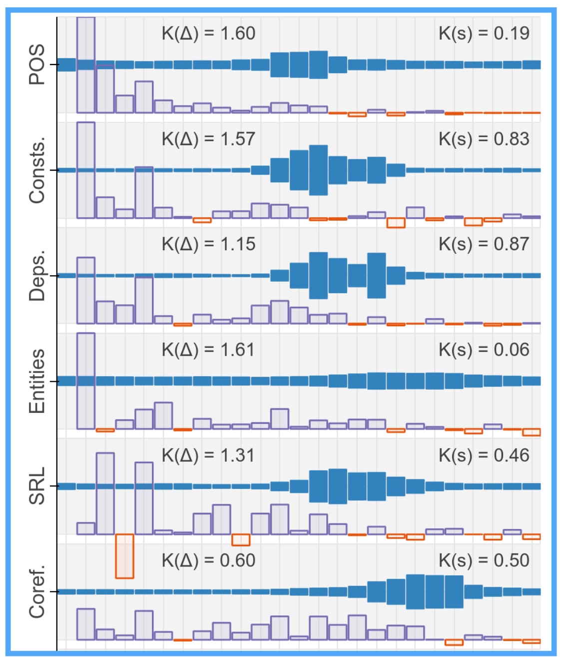
Rudinger et al. (2018)

Wealth of evidence that linguistic information is "there"

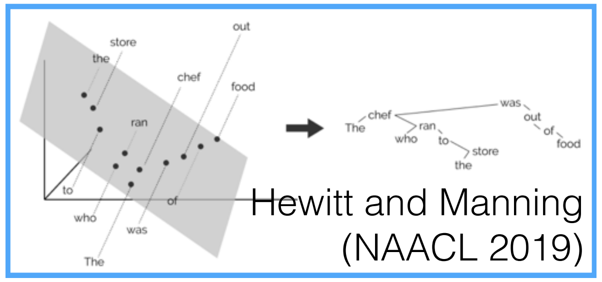
Past ~2 years:
What do deep LMs know about language?

What types of features **representations** encode? ("Probing Classifiers")

Do models **behave** like they are using these features? ("Challenge Tasks")



Tenney et al (ACL 2019)



Hewitt and Manning (NAACL 2019)

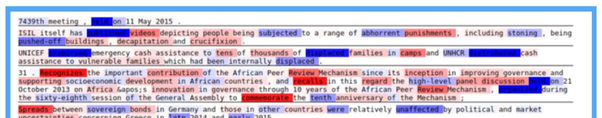
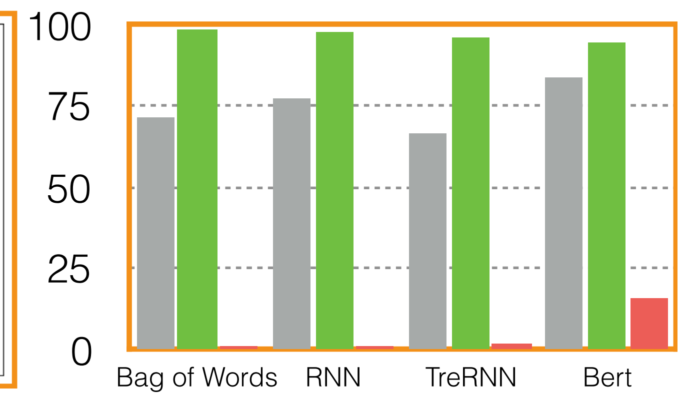


Figure 5: Visualization of a neuron from an English-Arabic model that activates on verb tense: negative/positive for past/present. Examples shown are the first 5 sentences in the test set.

Bau et al. (ICLR 2019)

Article: Super Bowl 50
Paragraph: *"Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV."*
Question: *"What is the name of the quarterback who was 38 in Super Bowl XXXIII?"*
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Jia and Liang (2017)



McCoy et al. (2019)

(1a) **The paramedic** performed CPR on the passenger even though she/he/they knew it was too late.

(2a) The paramedic performed CPR on **the passenger** even though she/he/they was/were already dead.

(1b) **The paramedic** performed CPR on someone even though she/he/they knew it was too late.

(2b) The paramedic performed CPR on **someone** even though she/he/they was/were already dead.

Rudinger et al. (2018)

Wealth of evidence that linguistic information is "there"

...but the model doesn't use it...

Linguistic features seem to be “there” after pretraining, but fine-tuned models don’t use them… why?

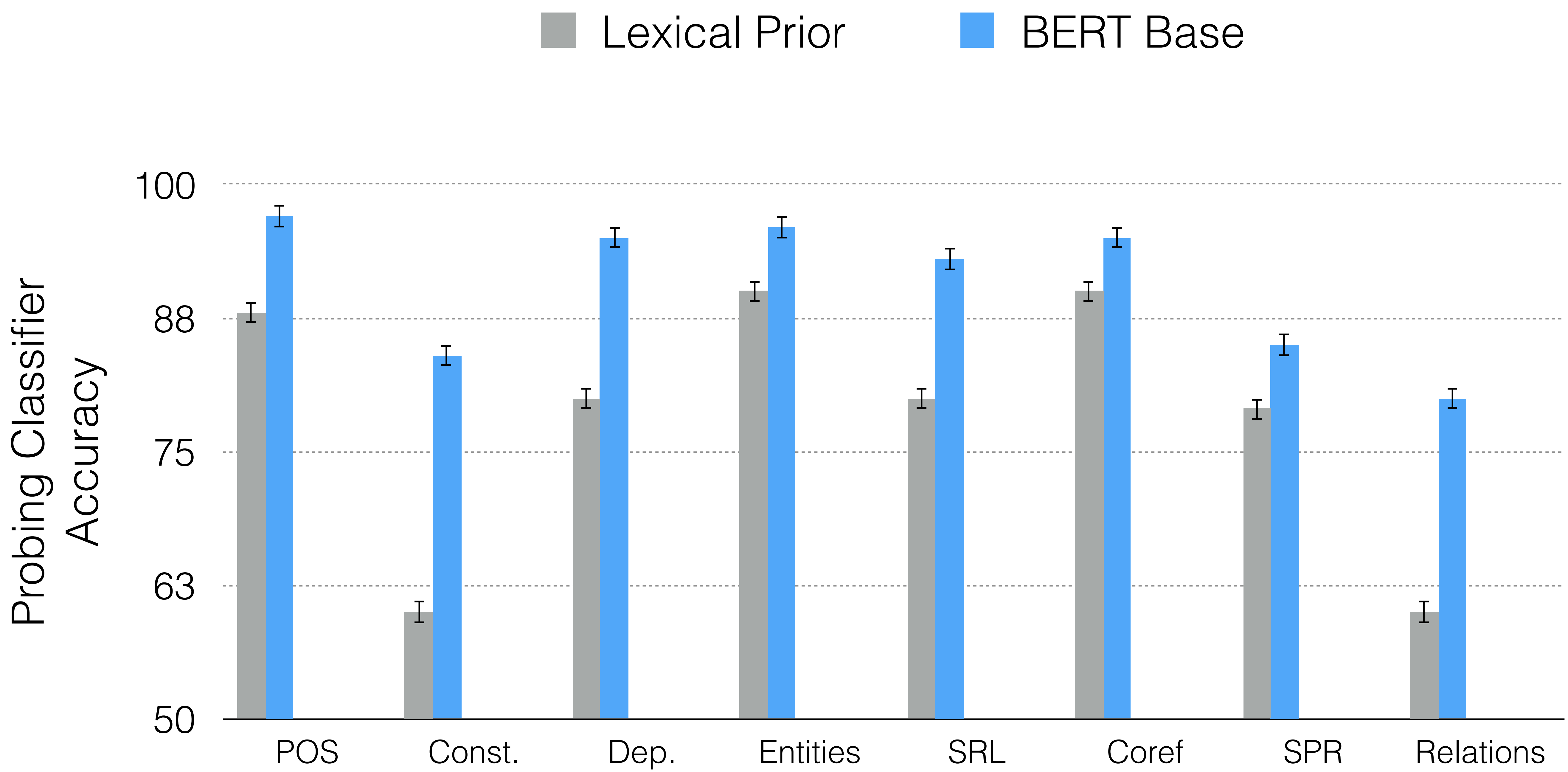
Linguistic features seem to be “there” after pretraining, but fine-tuned models don’t use them… why?

Maybe the features are erased during finetuning?

Are features lost during fine-tuning?

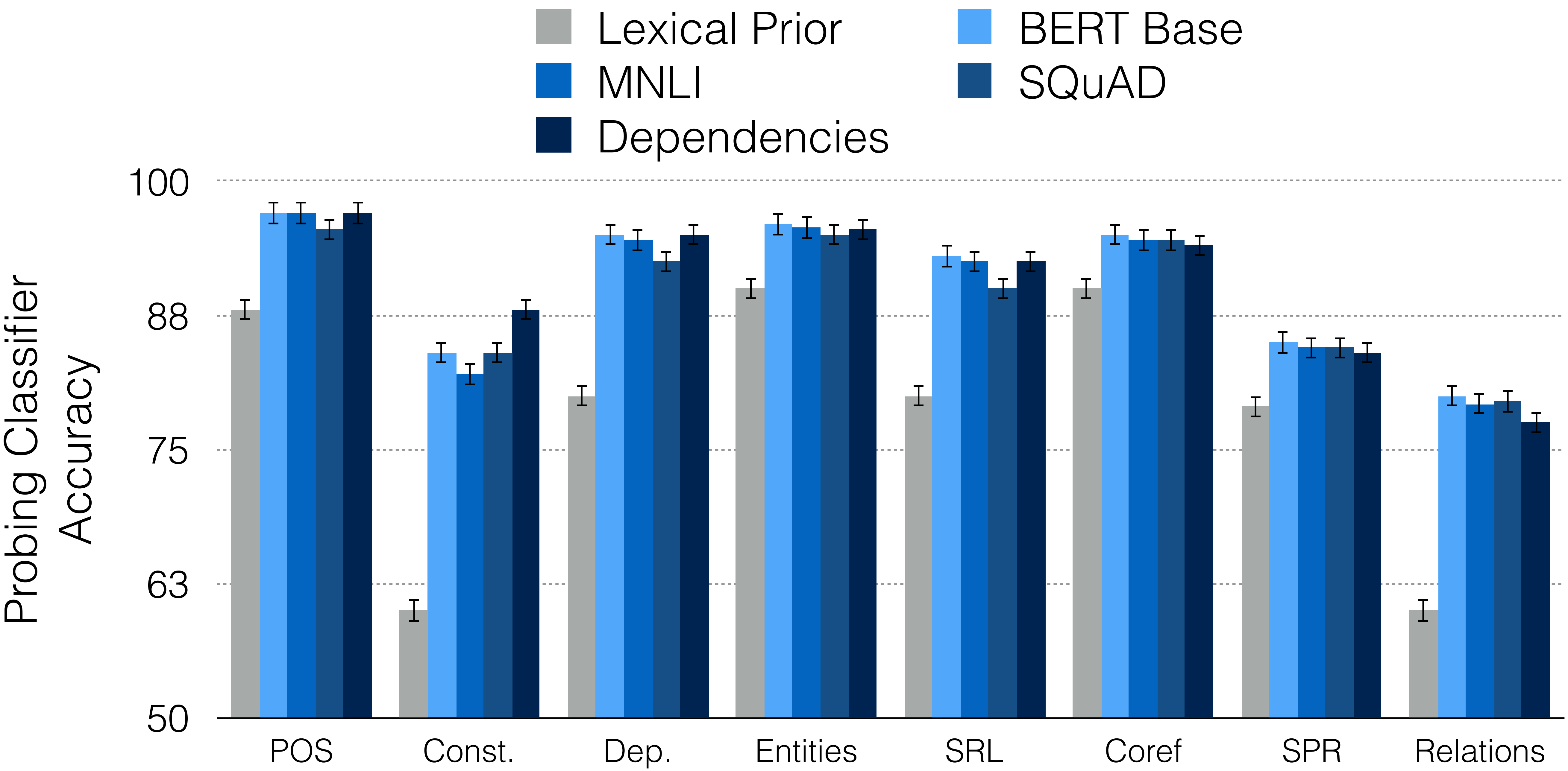
What Happens To BERT Embeddings During Fine-tuning?
Merchant, Rahimtoroghi, Pavlick, Tenney (2020).

Are features lost during fine-tuning?



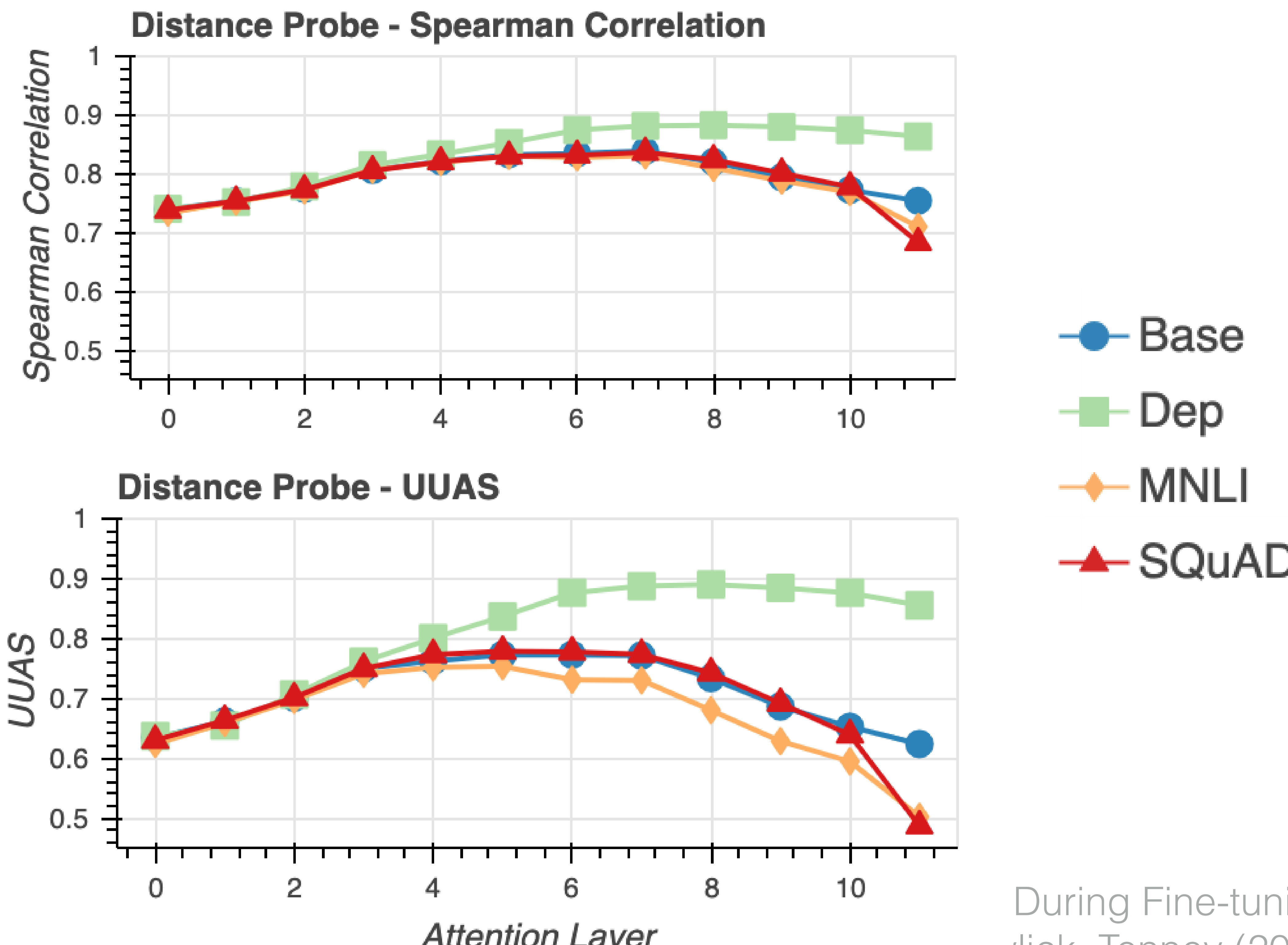
What Happens To BERT Embeddings During Fine-tuning?
Merchant, Rahimtoroghi, Pavlick, Tenney (2020).

Are features lost during fine-tuning?



What Happens To BERT Embeddings During Fine-tuning?
Merchant, Rahimtoroghi, Pavlick, Tenney (2020).

Are features lost during fine-tuning?



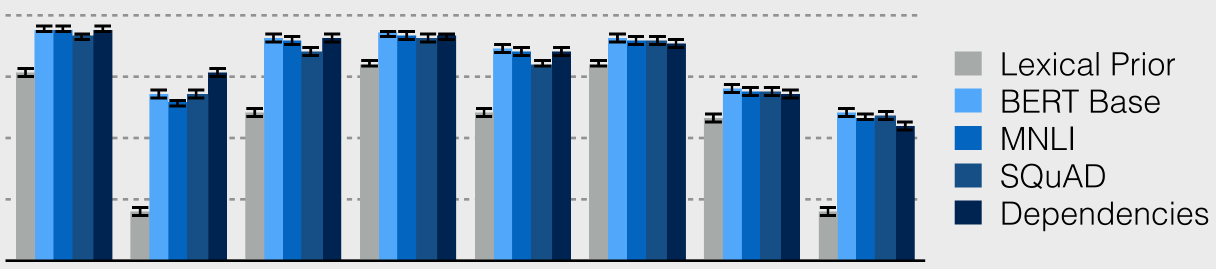
During Fine-tuning?
Merchant, Rahimtoroghi, Pavlick, Tenney (2020).

Linguistic features seem to be “there” after pretraining, but fine-tuned models don’t use them… why?

Maybe the features are erased during finetuning?

Linguistic features seem to be "there" after pretraining, but fine-tuned models don't use them… why?

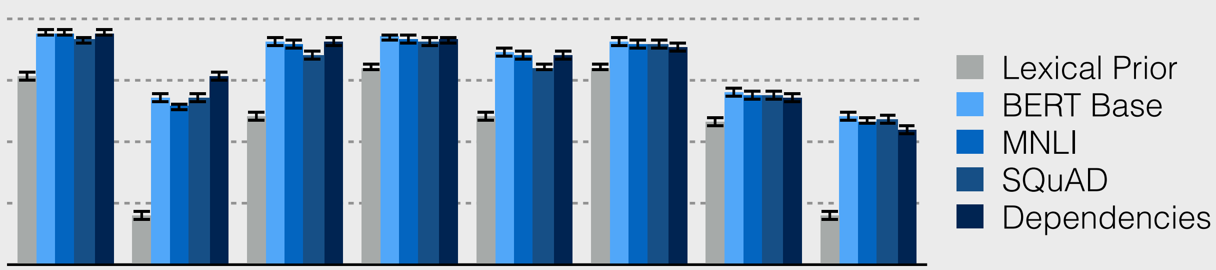
~~Maybe the features are erased during finetuning?~~



No obvious drop in probing accuracy after fine-tuning.

Linguistic features seem to be “there” after pretraining, but fine-tuned models don’t use them… why?

~~Maybe the features are erased during finetuning?~~

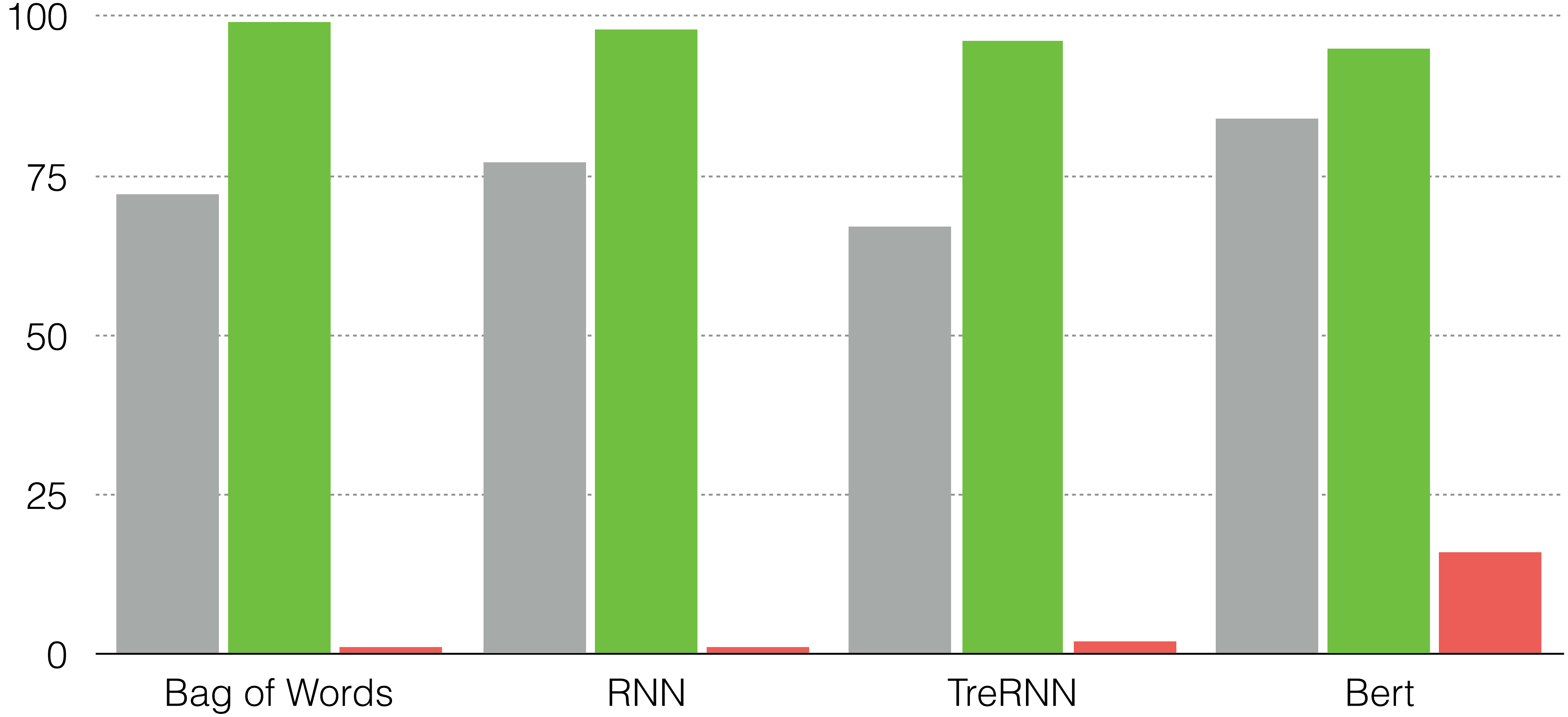


No obvious drop in probing accuracy after fine-tuning.

Maybe there just isn’t enough signal in training?

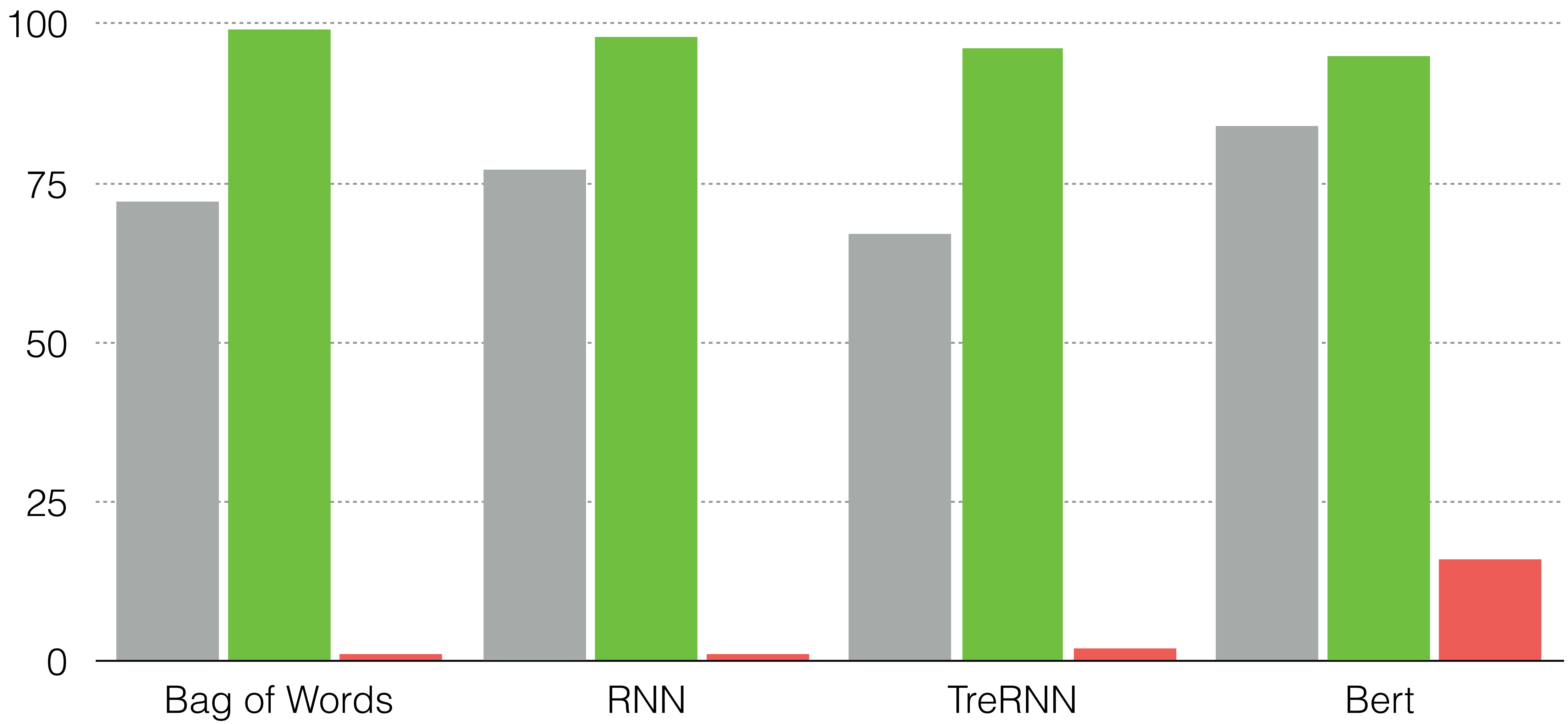
Blame it on the training data?

Blame it on the training data?



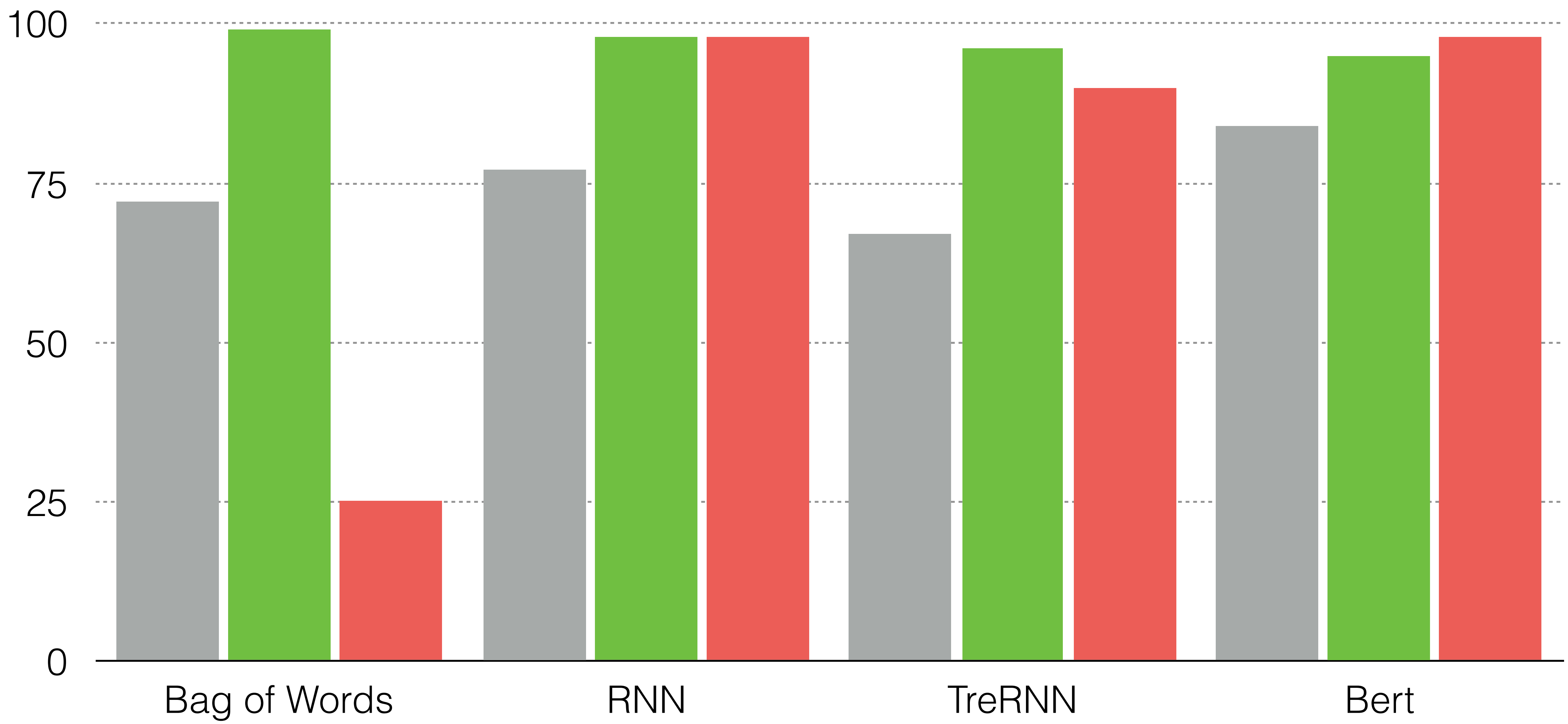
Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference.
McCoy, Pavlick, and Linzen (2019)

Blame it on the training



Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference.
McCoy, Pavlick, and Linzen (2019)

Blame it on the training



Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference.
McCoy, Pavlick, and Linzen (2019)

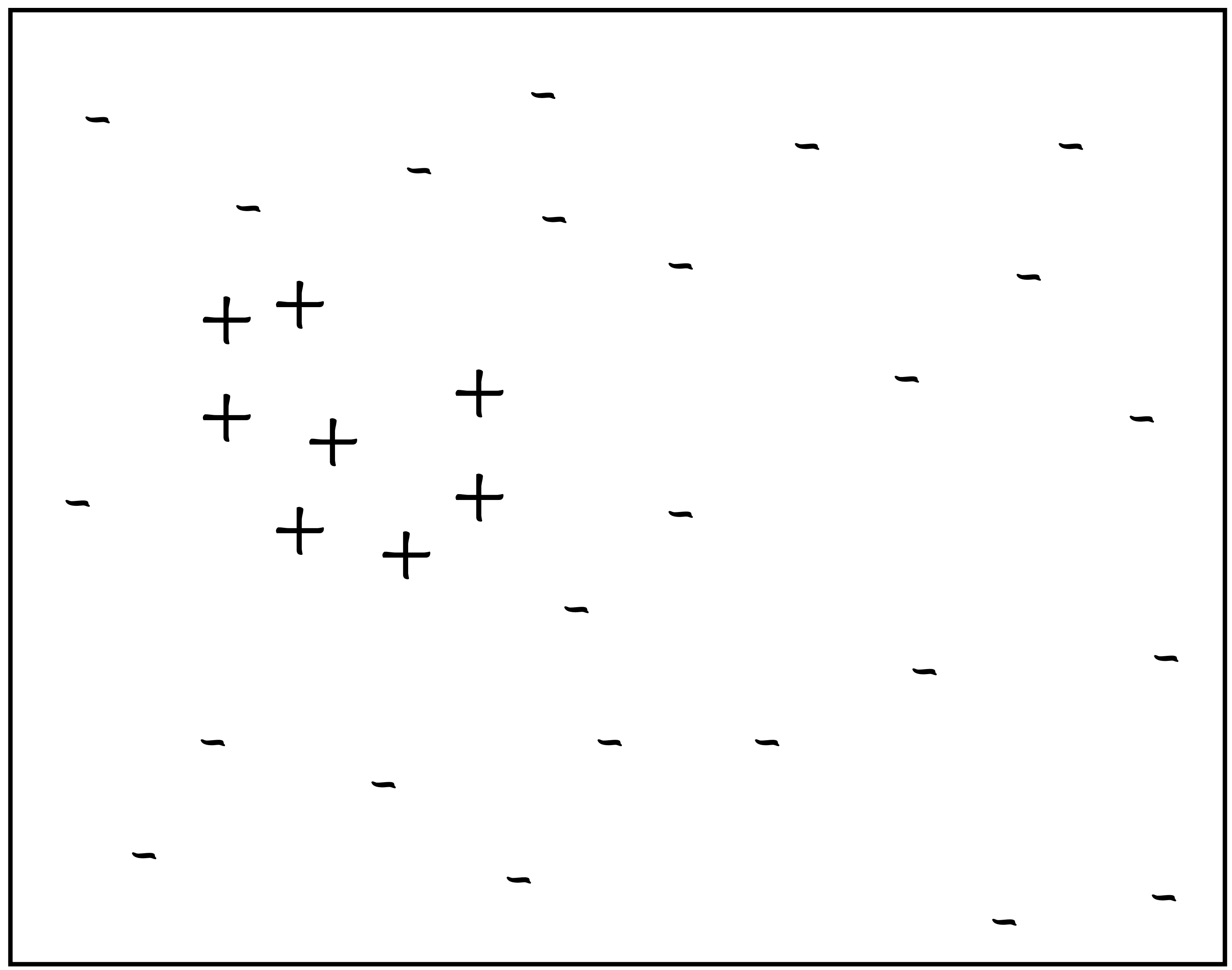
Blame it on the training

Targeted/Adversarial/
Counterfactual Data Augmentation

Zhao et al. (2018)
Elkahky et al. (2018)
Zmigrod et al. (2019)
McCoy et al. (2019)
Lu et al. (2019)
Hall Maudslay et al. (2019)
...

Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference.
McCoy, Pavlick, and Linzen (2019)

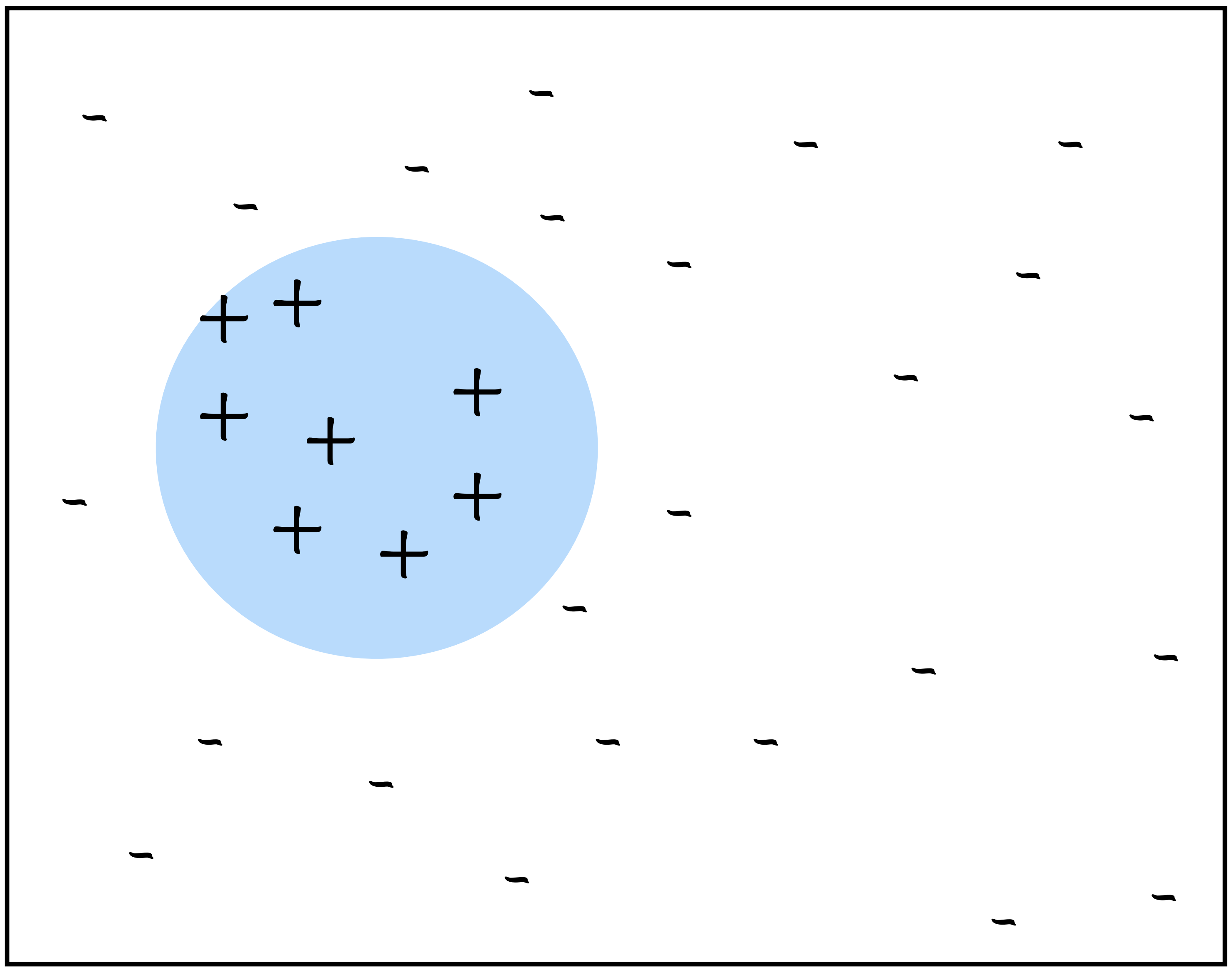
General Set Up



Information-Theoretic Probing Explains Reliance on Spurious Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

General Set Up

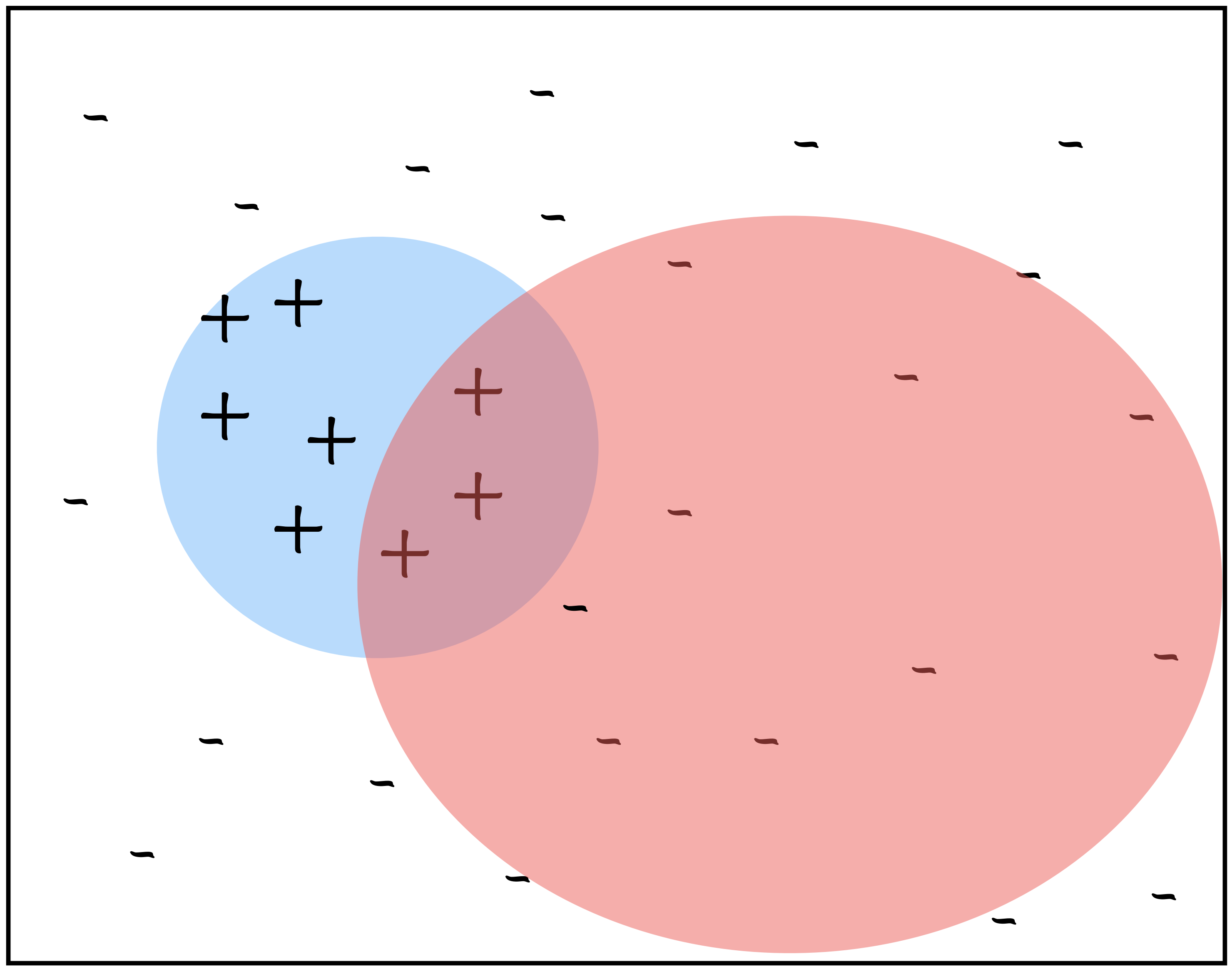
"Target" feature perfectly predicts label



Information-Theoretic Probing Explains Reliance on Spurious Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

General Set Up

“Target” feature perfectly predicts label

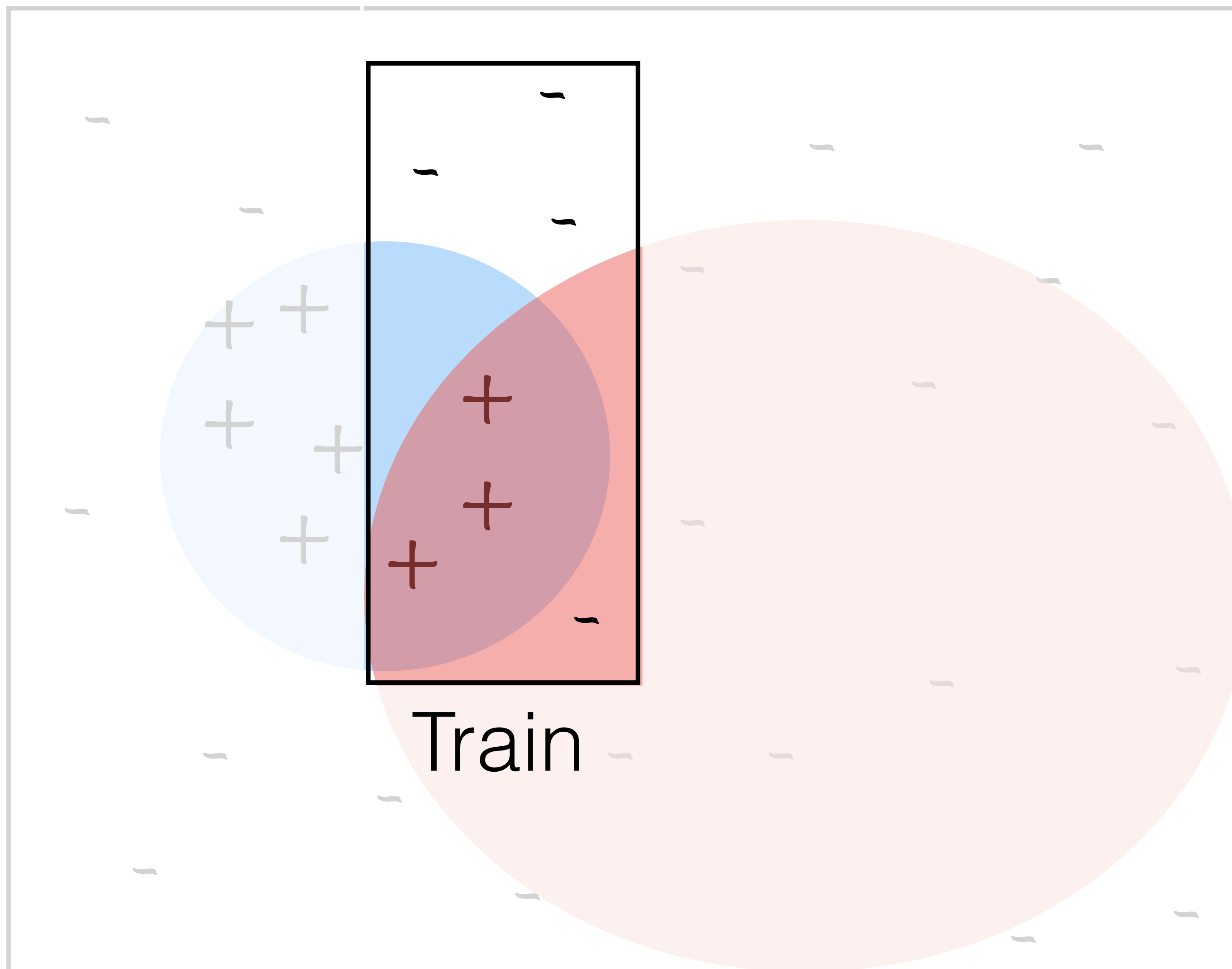


“Spurious” feature

Information-Theoretic Probing Explains Reliance on Spurious Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

General Set Up

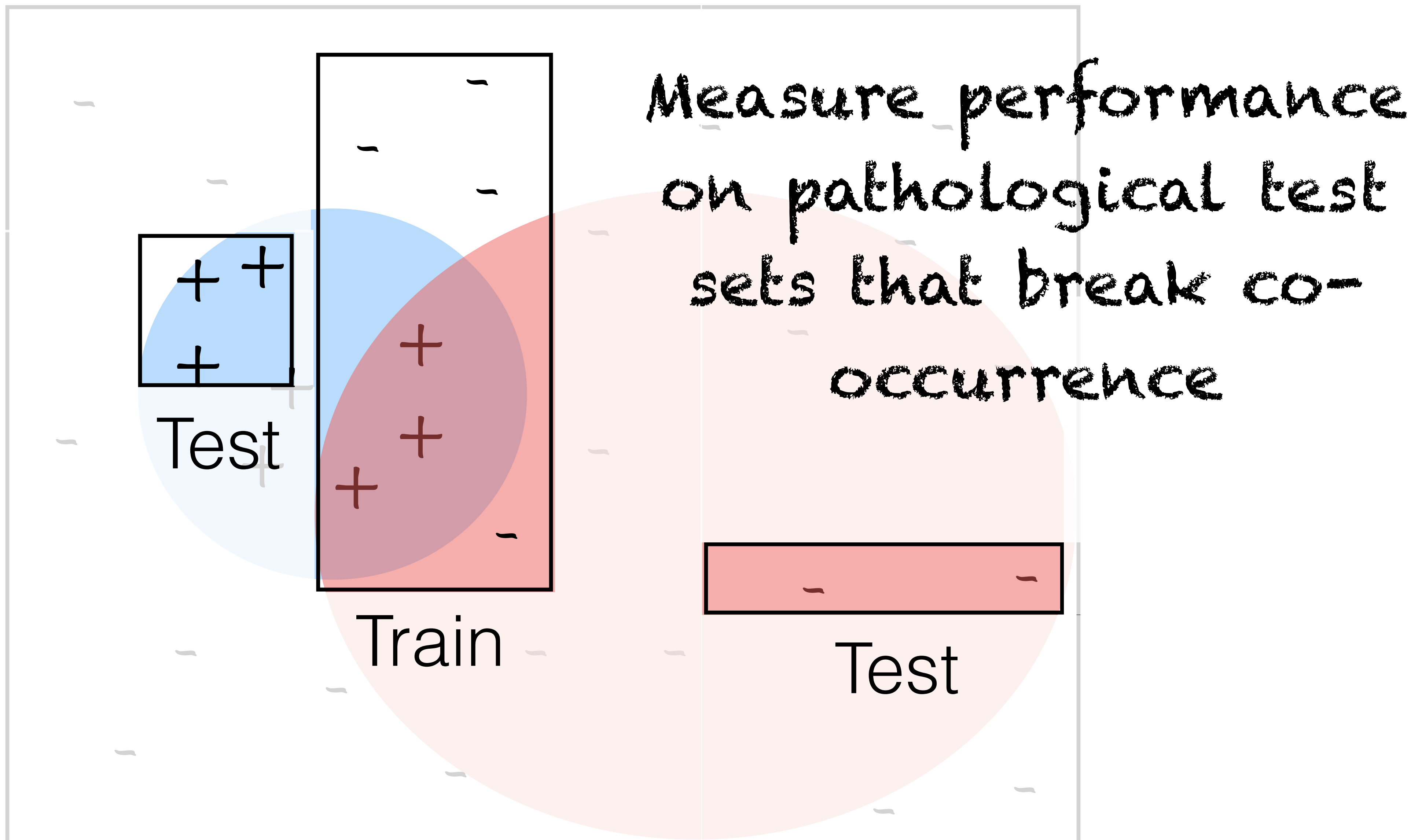
"Target" feature perfectly predicts label



"Spurious" feature which happens to co-occur with target in training sample

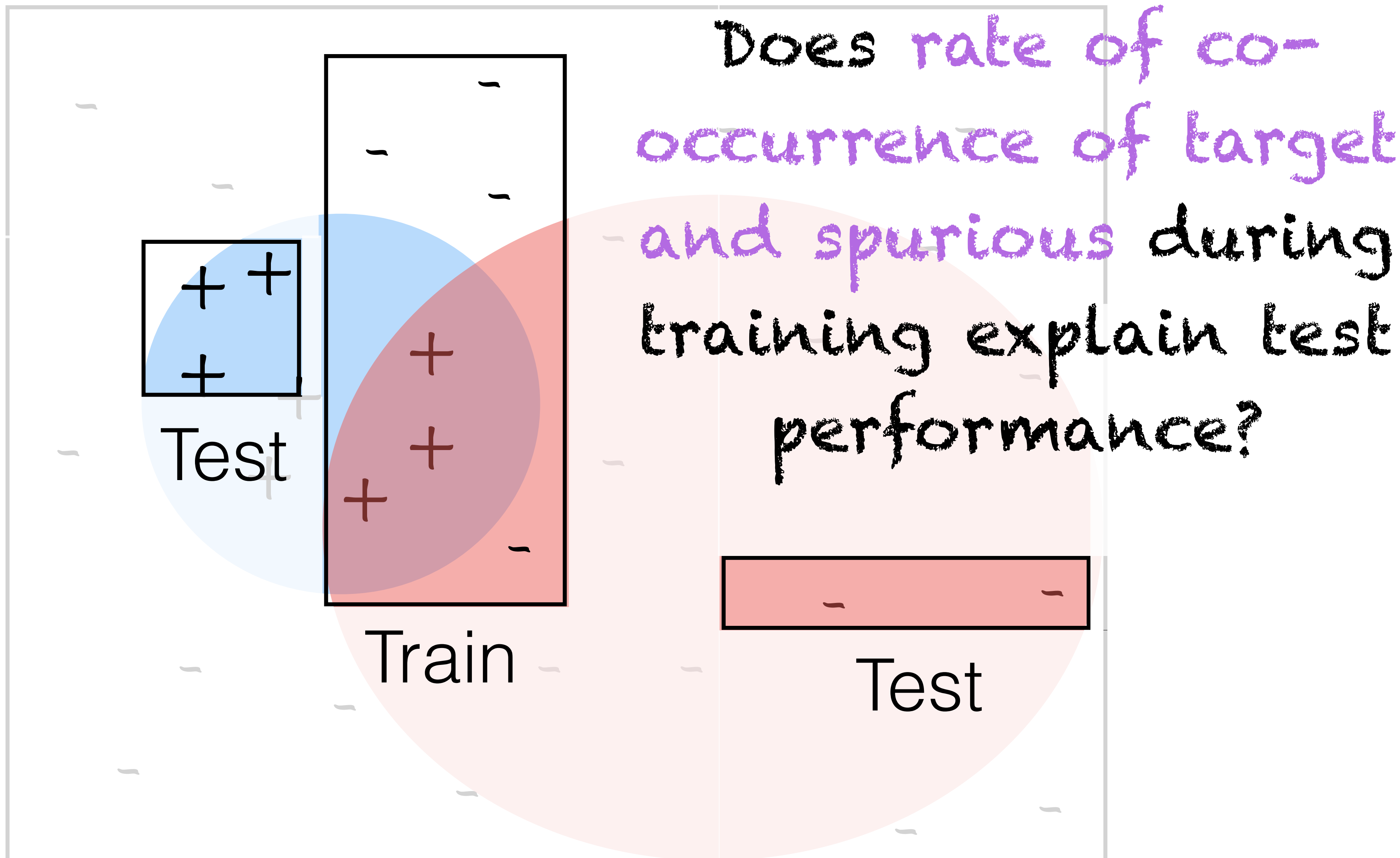
Information-Theoretic Probing Explains Reliance on Spurious Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

General Set Up



Information-Theoretic Probing Explains Reliance on Spurious Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

General Set Up



Information-Theoretic Probing Explains Reliance on Spurious Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

Toy Sentence Classification Task

Name	Target	Spurious	Example
`contains-1`	a '1' occurs in the sequence	a '2' occurs in the sequence	2 4 11 1 4
`prefix-duplicate`	sequence begins with a duplicate	a '2' occurs in the sequence	5 5 11 12 2
`adjacent-duplicate`	duplicate occurs somewhere in the sequence	a '2' occurs in the sequence	11 12 3 3 2
`first-last`	first symbol and last symbol are the same	a '2' occurs in the sequence	7 2 11 12 7

Information-Theoretic Probing Explains Reliance on Spurious Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

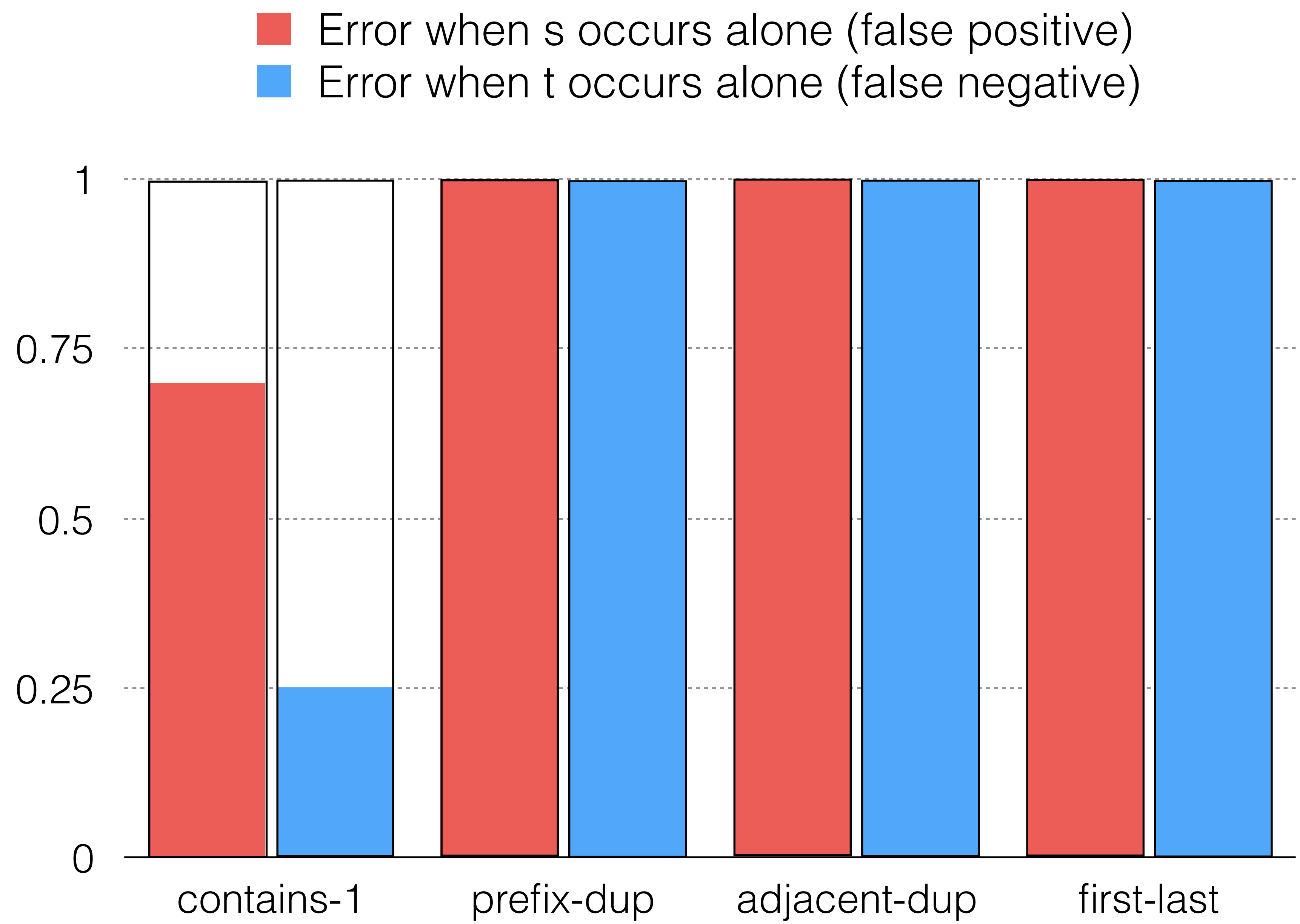
Out-of-Distribution Test Error

Perfect co-occurrence between spurious and target

Information-Theoretic Probing Explains Reliance on Spurious Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

Out-of-Distribution Test Error

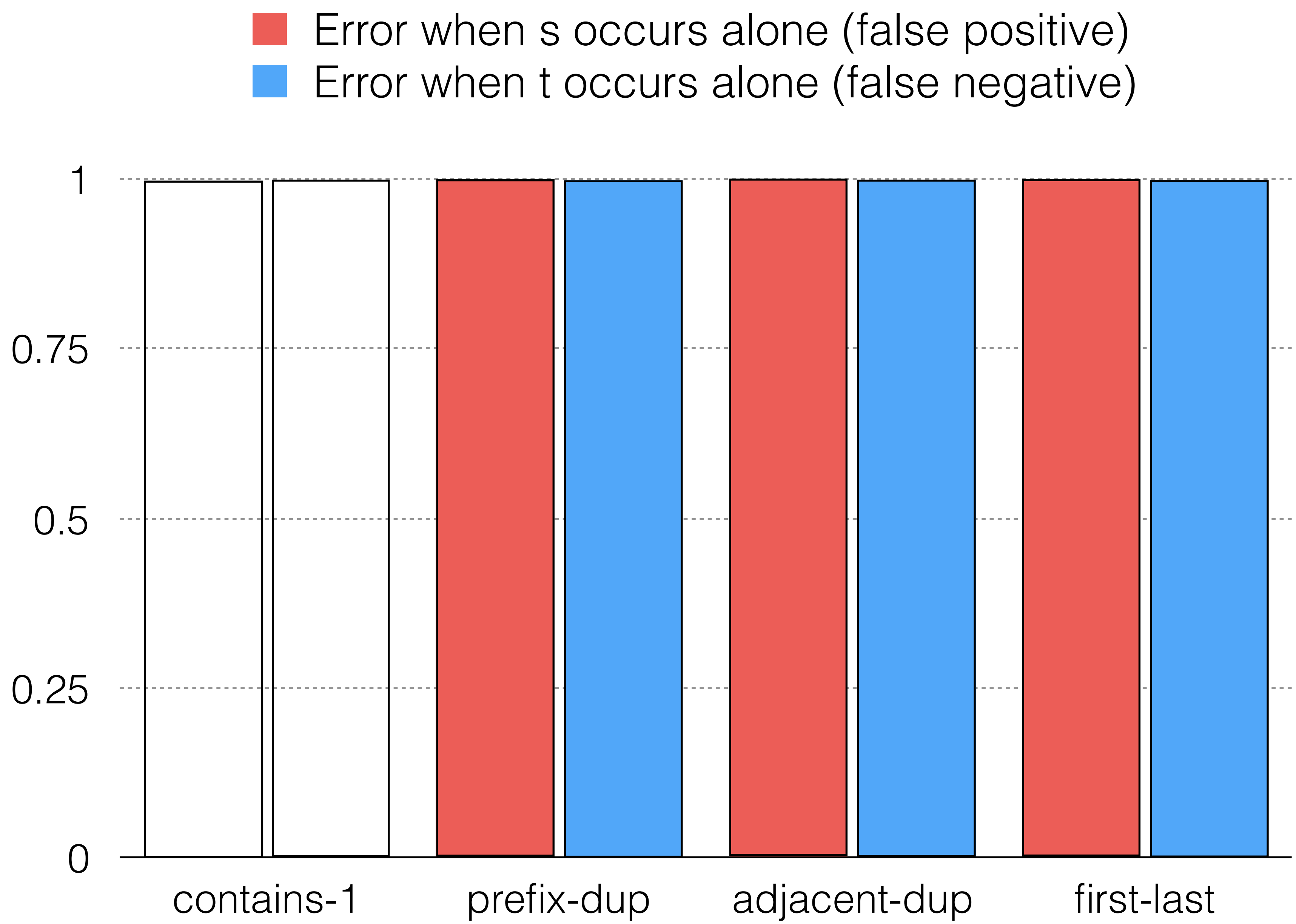
Perfect co-occurrence between spurious and target



Information-Theoretic Probing Explains Reliance on Spurious Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

Out-of-Distribution Test Error

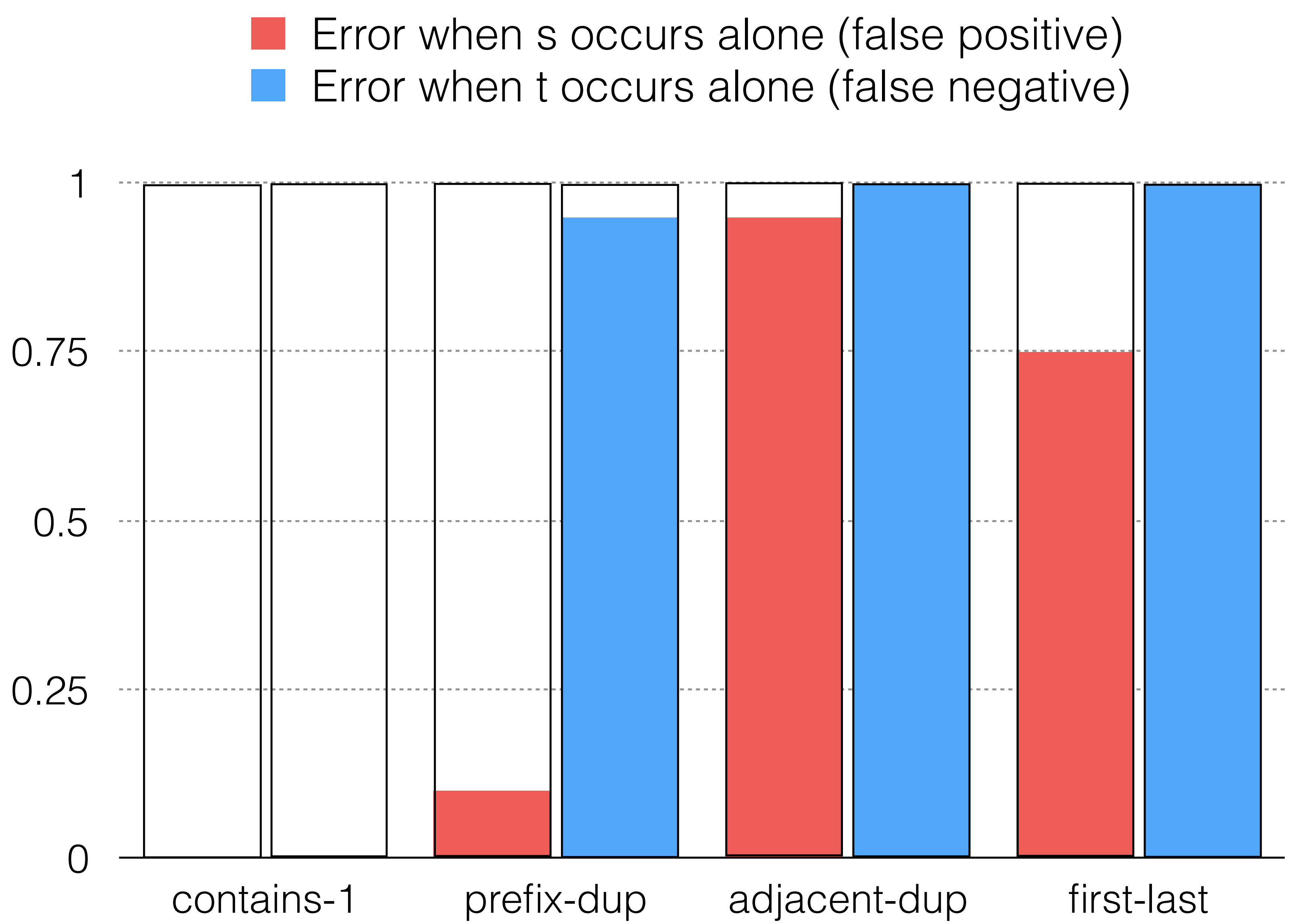
Spurious occurs without target in **0.1%** of training examples



Information-Theoretic Probing Explains Reliance on Spurious Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

Out-of-Distribution Test Error

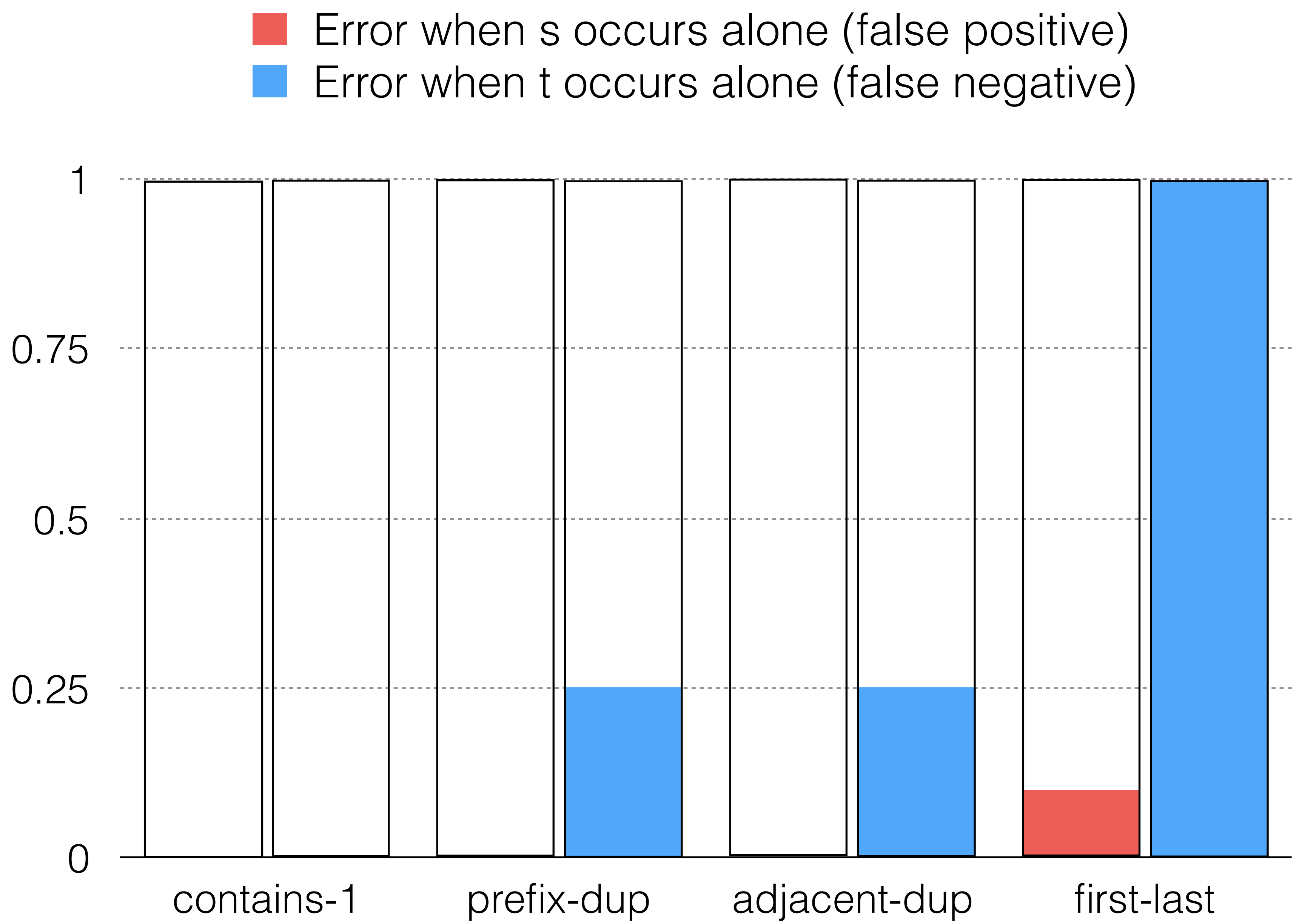
Spurious occurs without target in **10%** of training examples



Information-Theoretic Probing Explains Reliance on Spurious Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

Out-of-Distribution Test Error

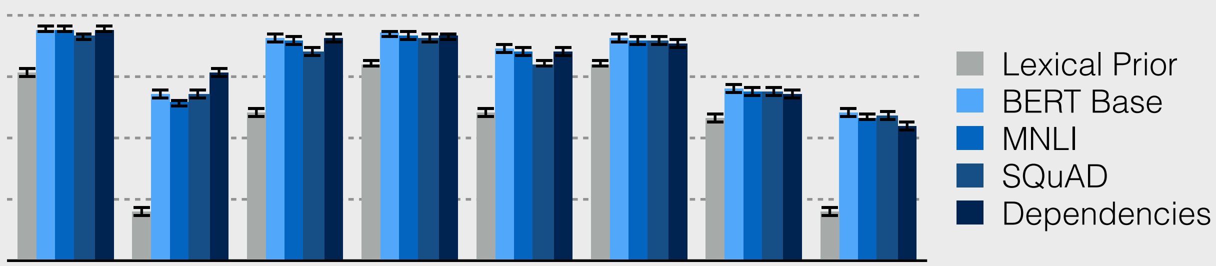
Spurious occurs without target in **50%** of training examples



Information-Theoretic Probing Explains Reliance on Spurious Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

Linguistic features seem to be "there" after pretraining, but fine-tuned models don't use them… why?

~~Maybe the features are erased during finetuning?~~

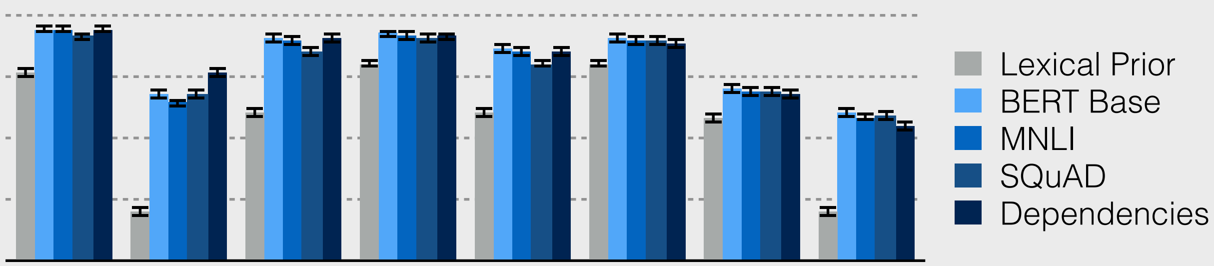


No obvious drop in probing accuracy after fine-tuning.

Maybe there just isn't enough signal in training?

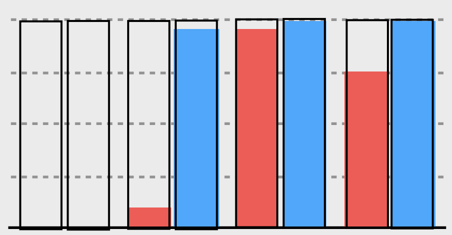
Linguistic features seem to be “there” after pretraining, but fine-tuned models don’t use them… why?

~~Maybe the features are erased during finetuning?~~



No obvious drop in probing accuracy after fine-tuning.

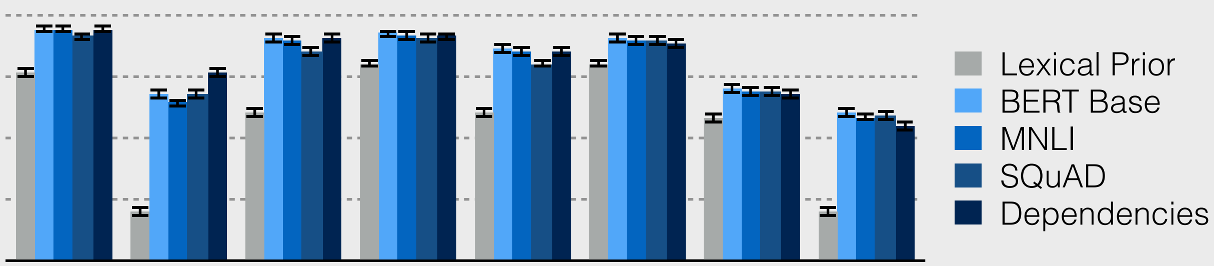
~~Maybe there just isn’t enough signal in training?~~



Different features behave differently given the same training data.

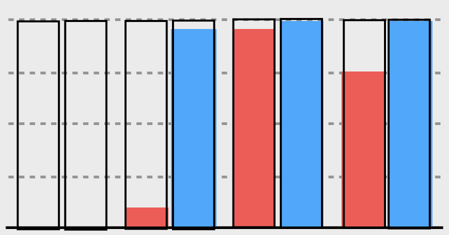
Linguistic features seem to be "there" after pretraining, but fine-tuned models don't use them… why?

~~Maybe the features are erased during finetuning?~~



No obvious drop in probing accuracy after fine-tuning.

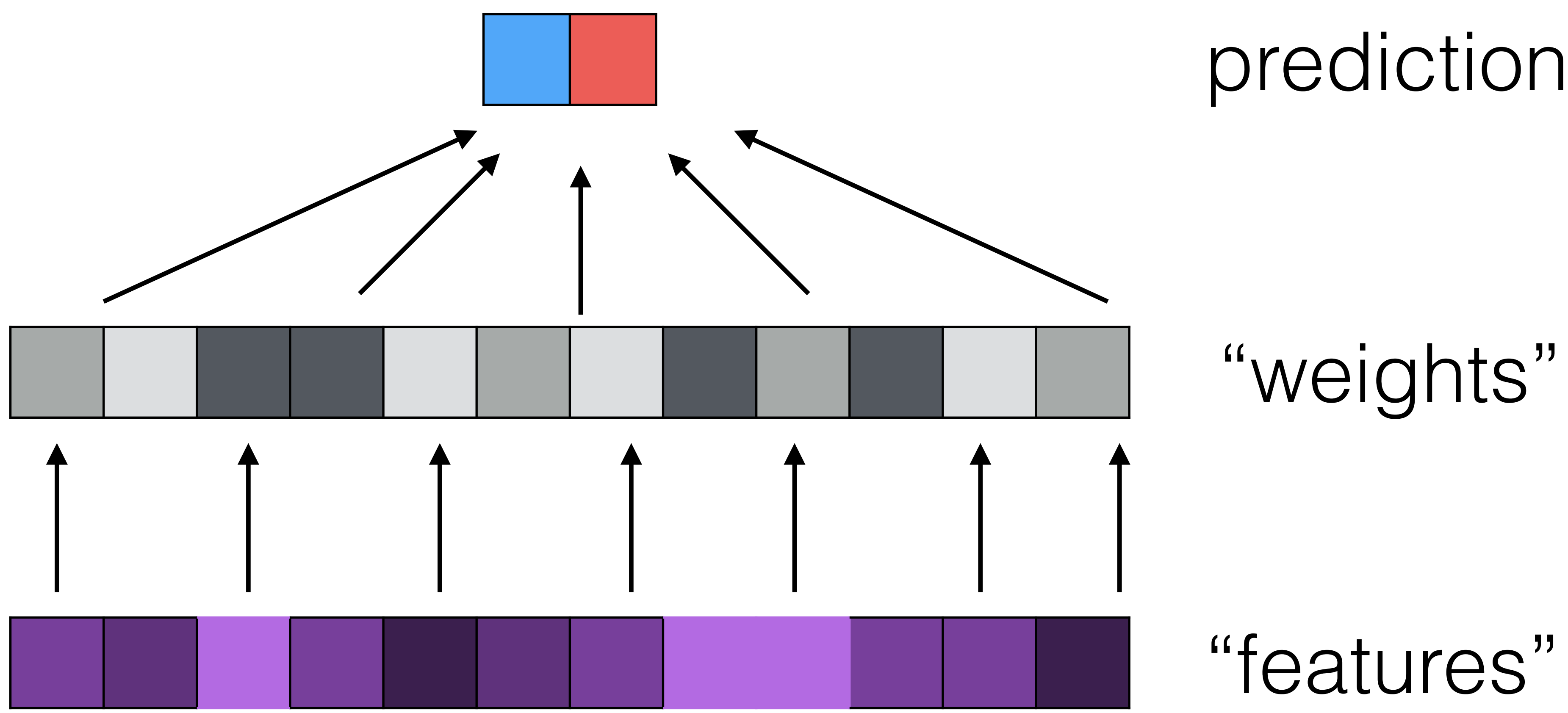
~~Maybe there just isn't enough signal in training?~~



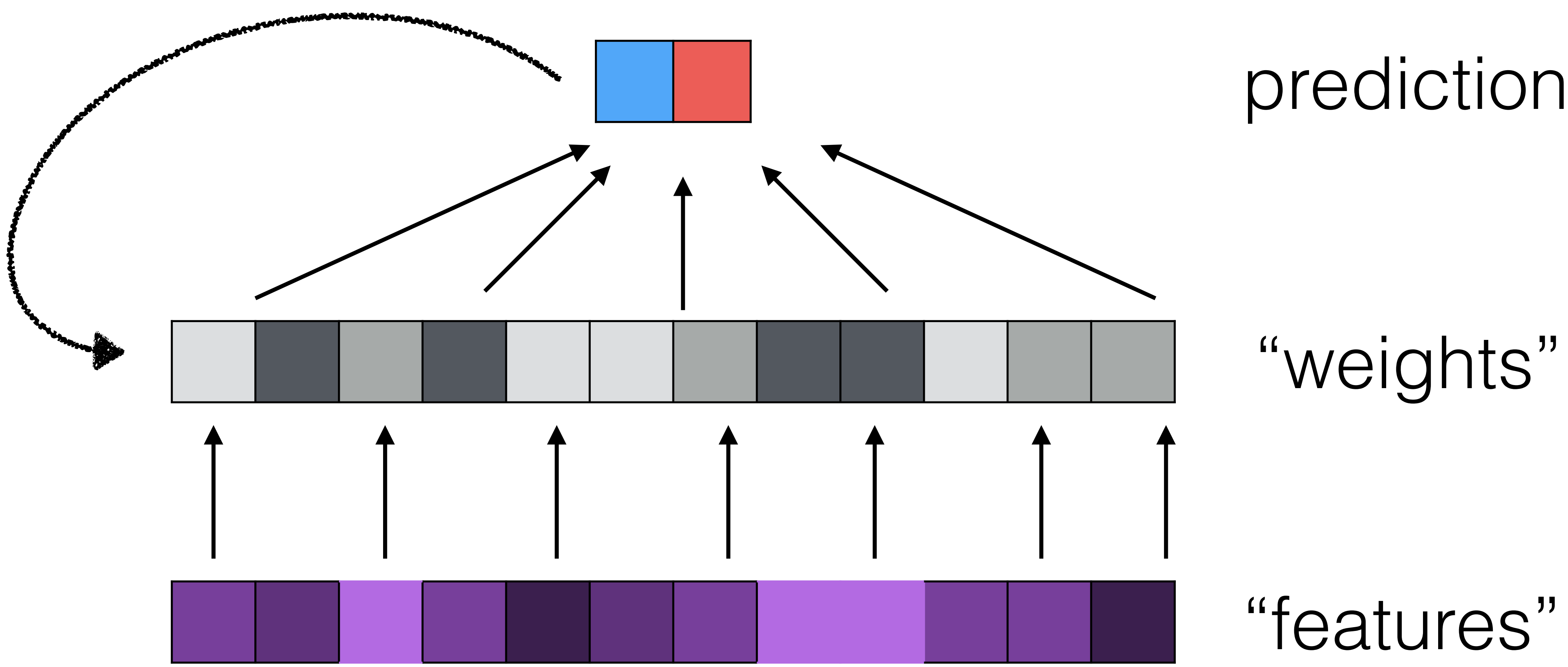
Different features behave differently given the same training data.

Maybe it's not just a matter of features being "there" or "not there"…?

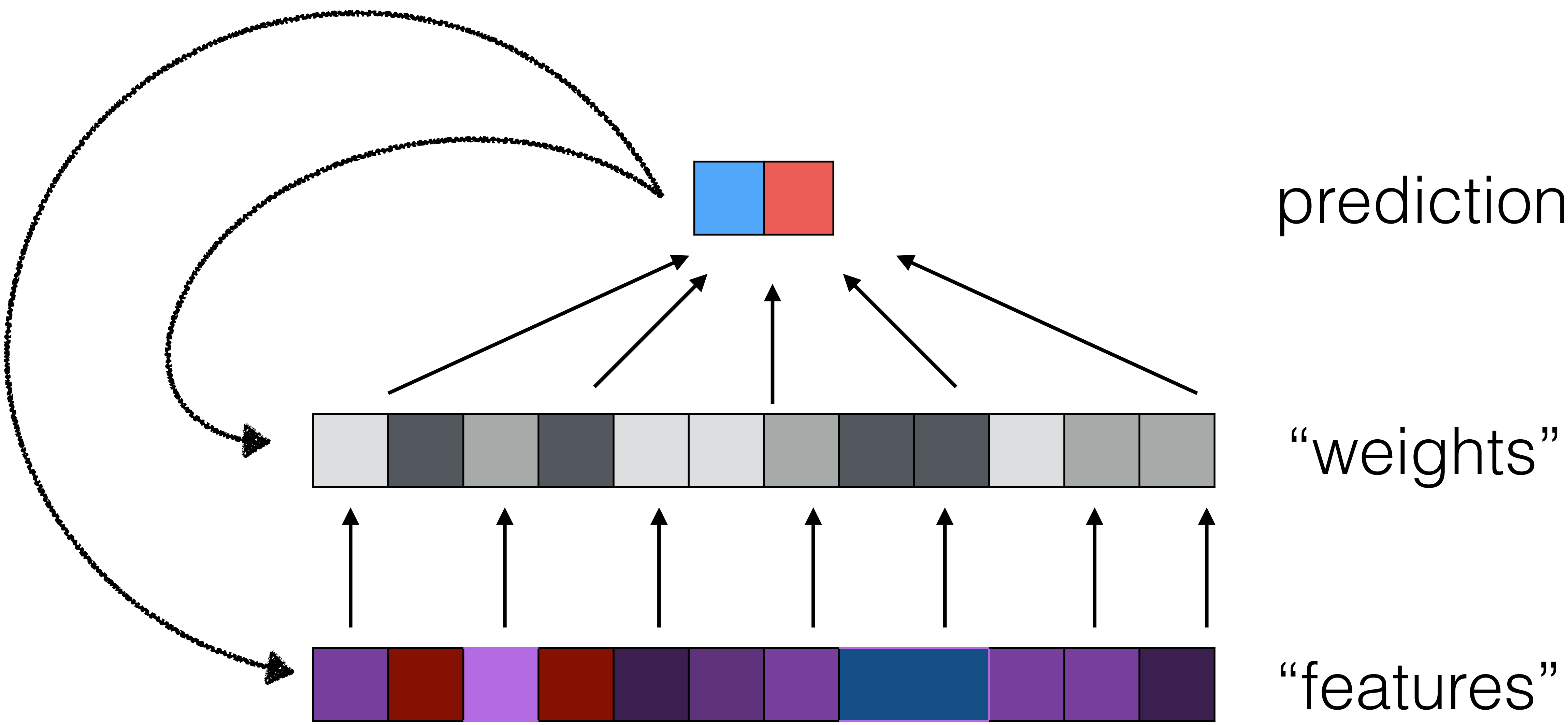
Learning Features vs. Learning Classifiers



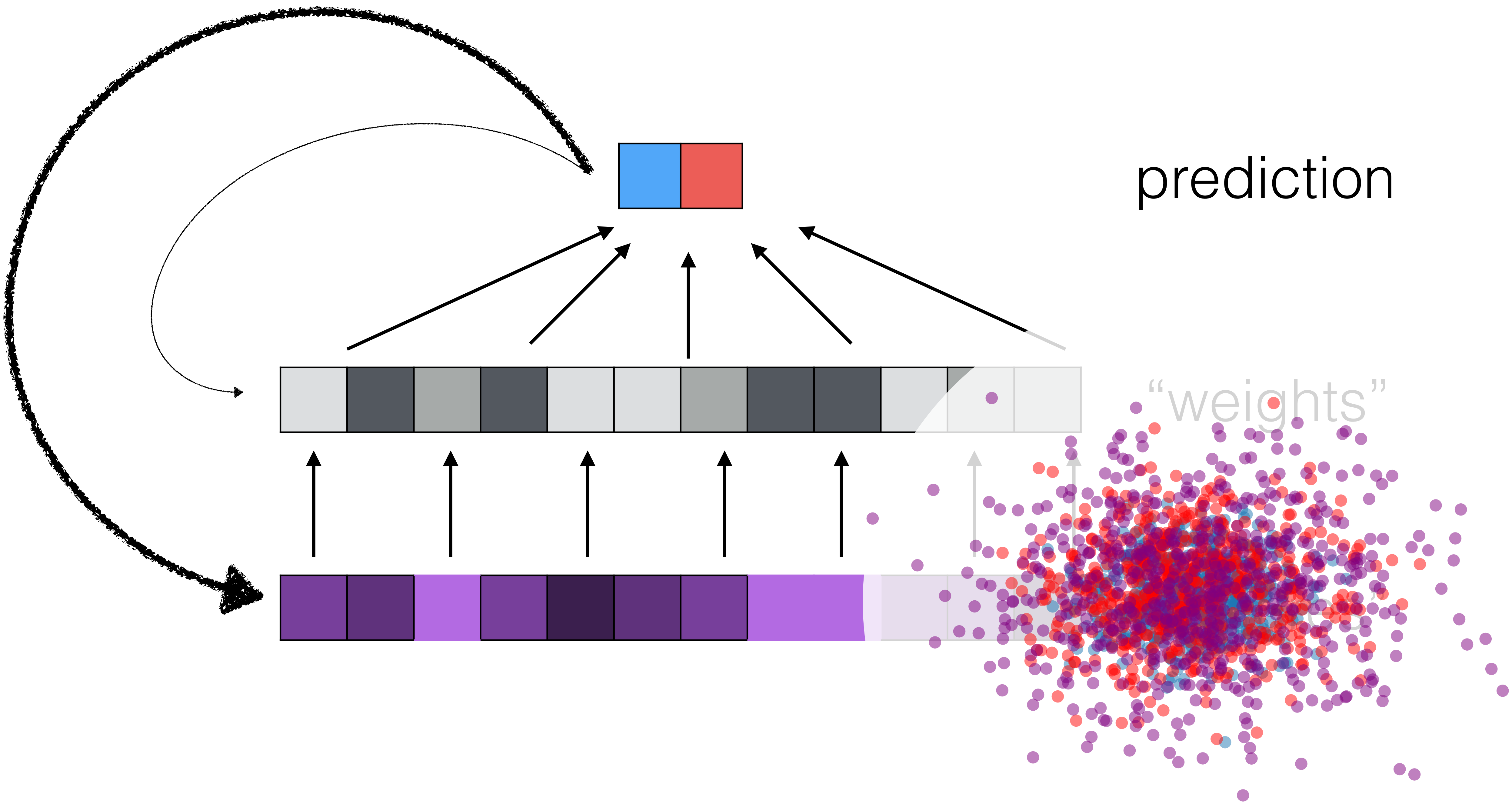
Learning Features vs. Learning Classifiers



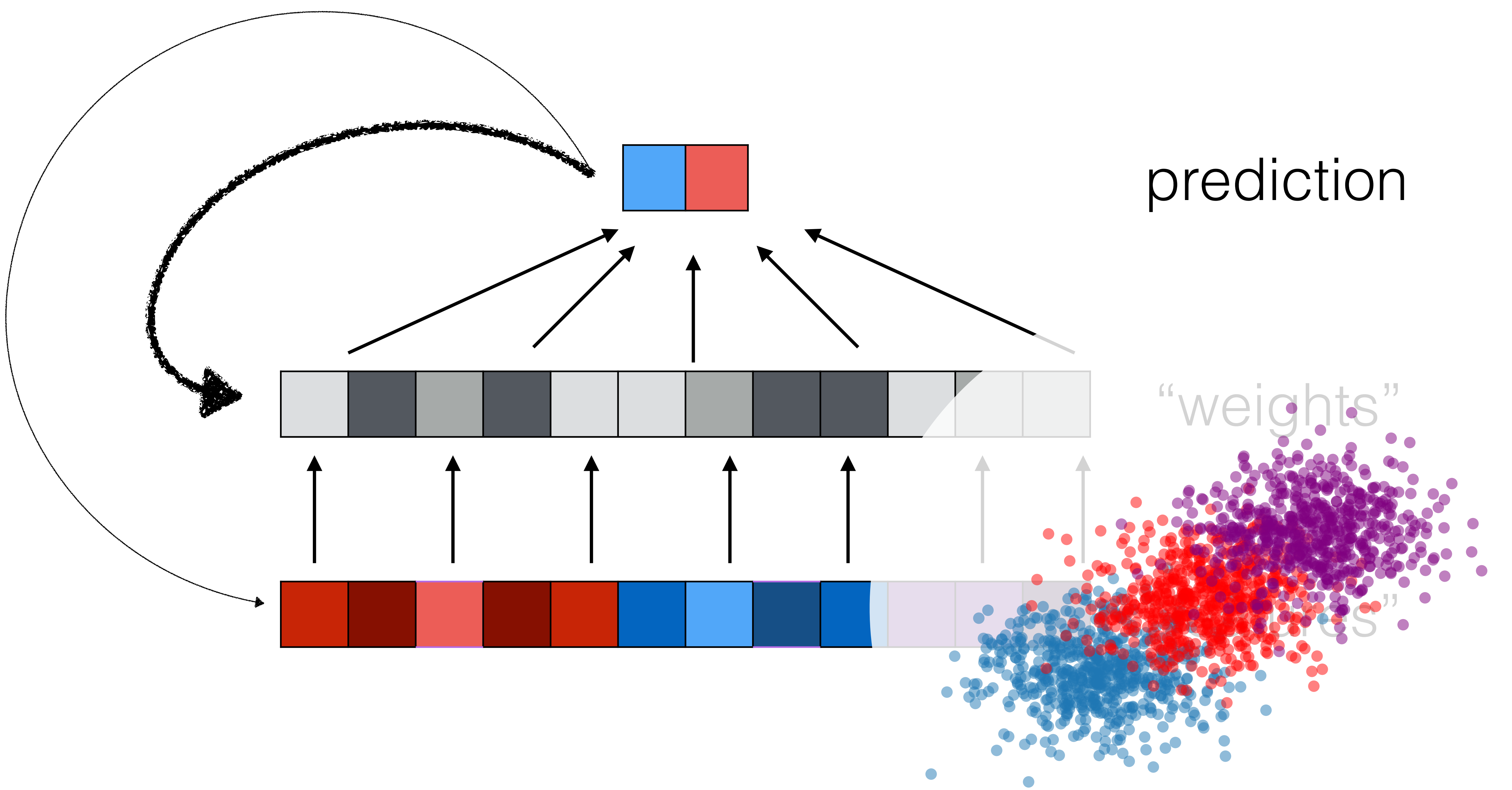
Learning Features vs. Learning Classifiers



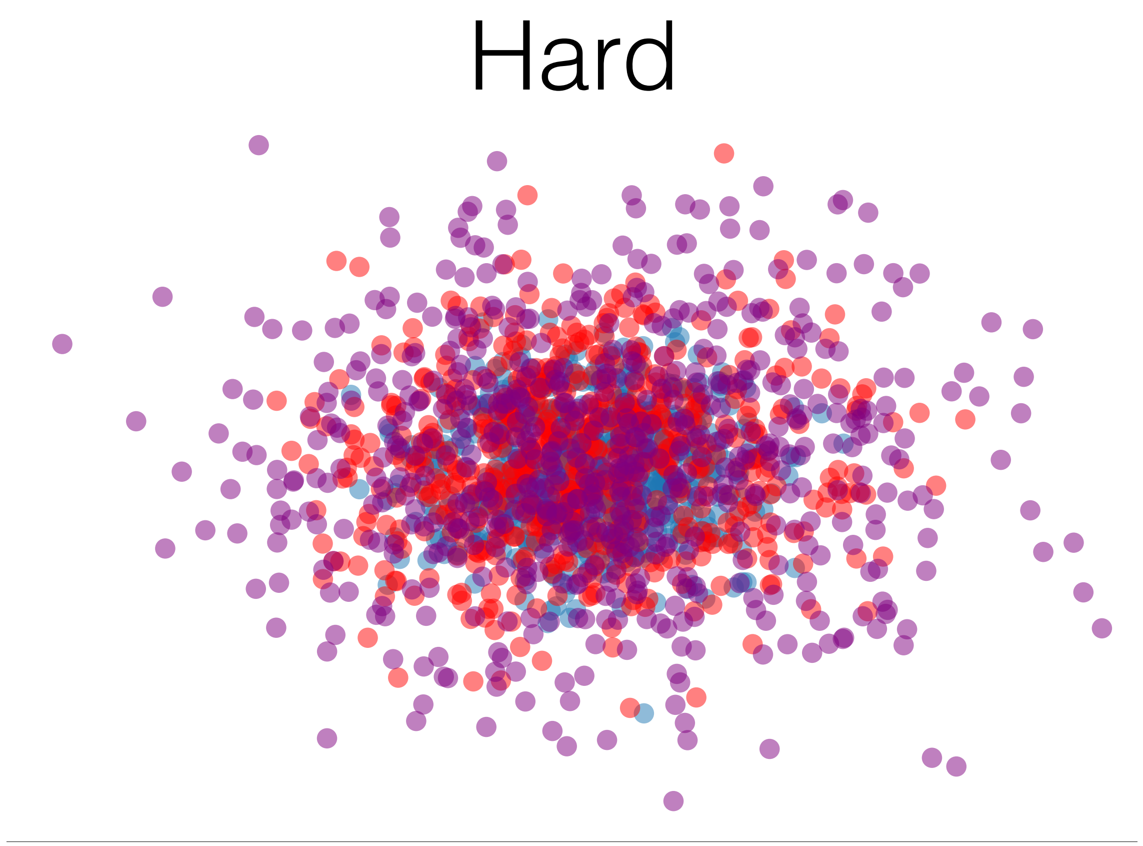
Learning Features vs. Learning Classifiers

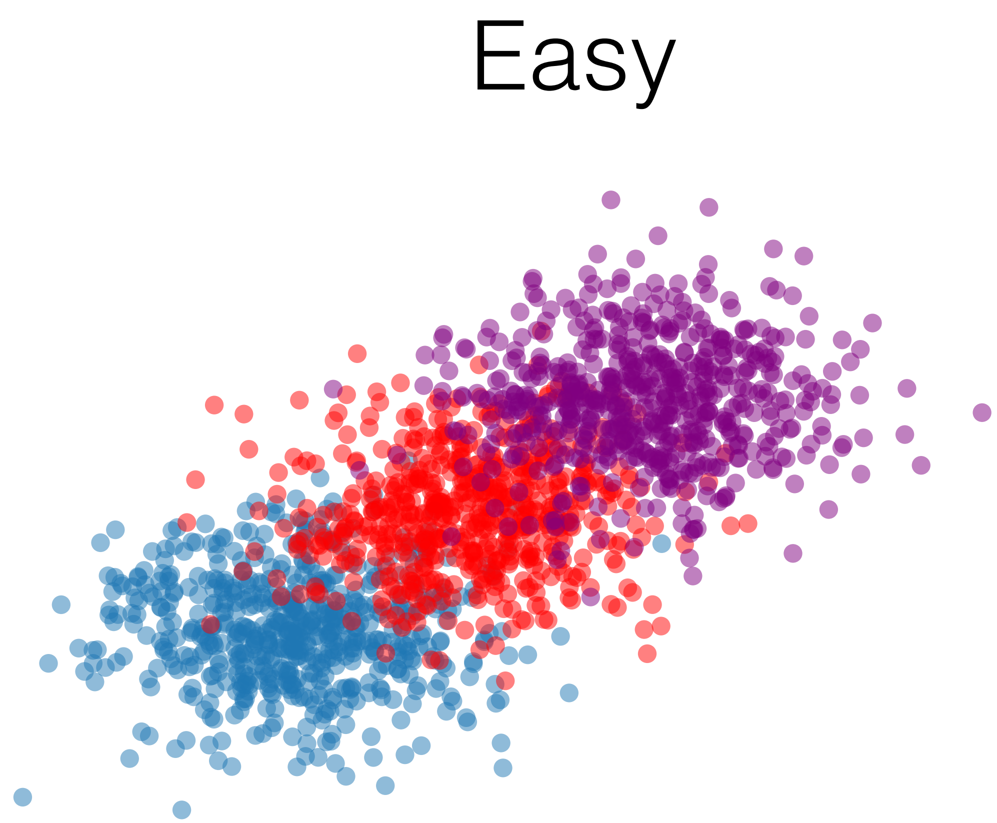


Learning Features vs. Learning Classifiers



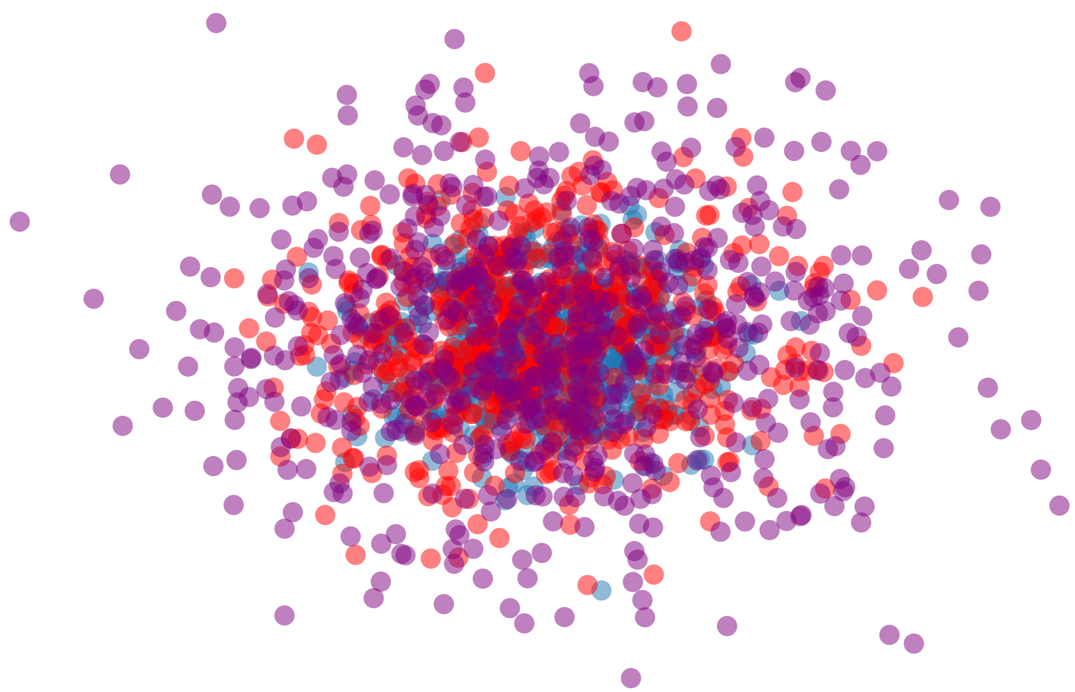
Features differ in how "hard" they are to extract



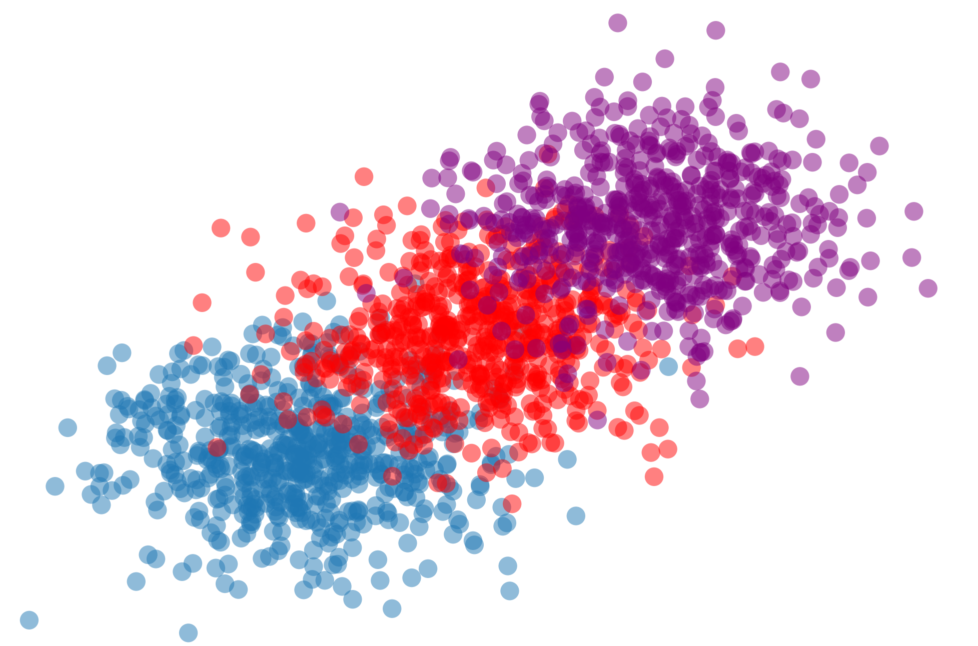


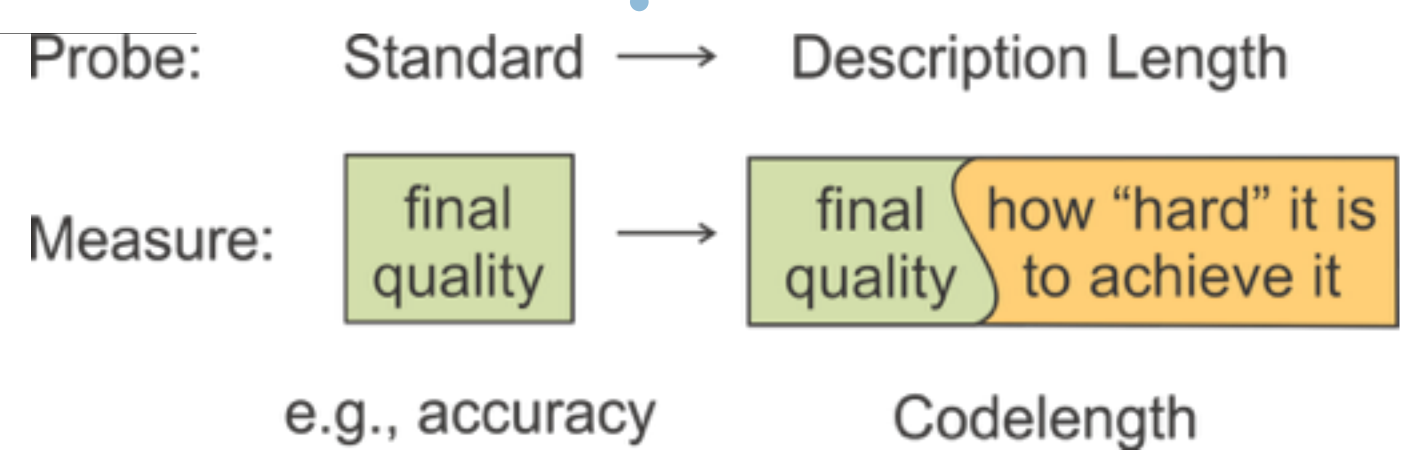
Features differ in how "hard" they are to extract

Hard



Easy

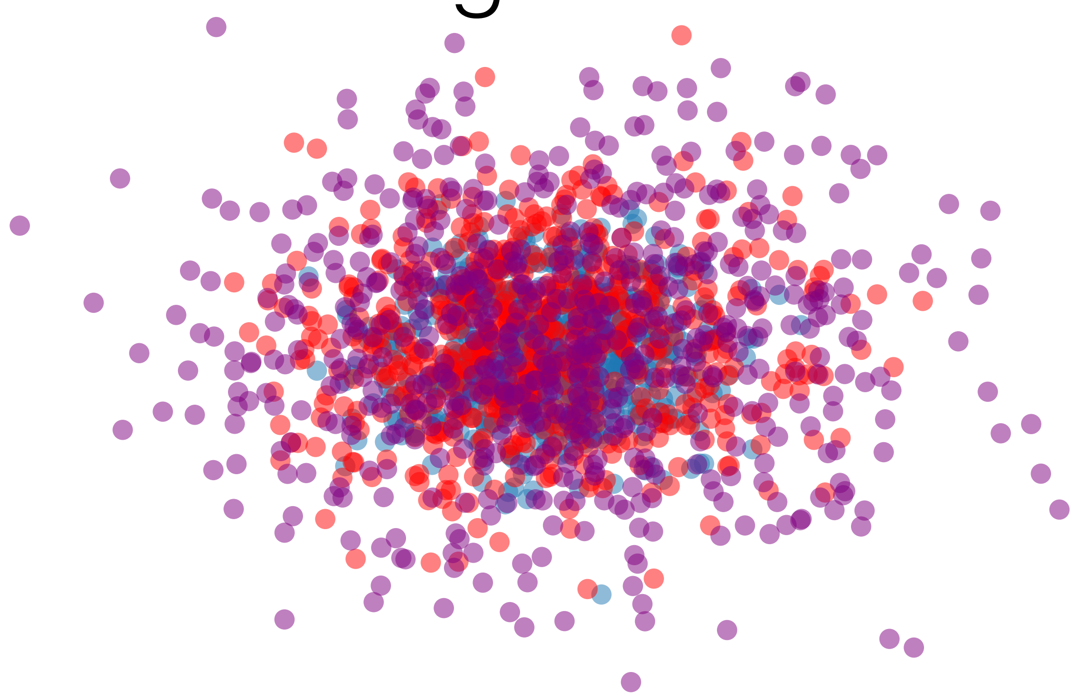




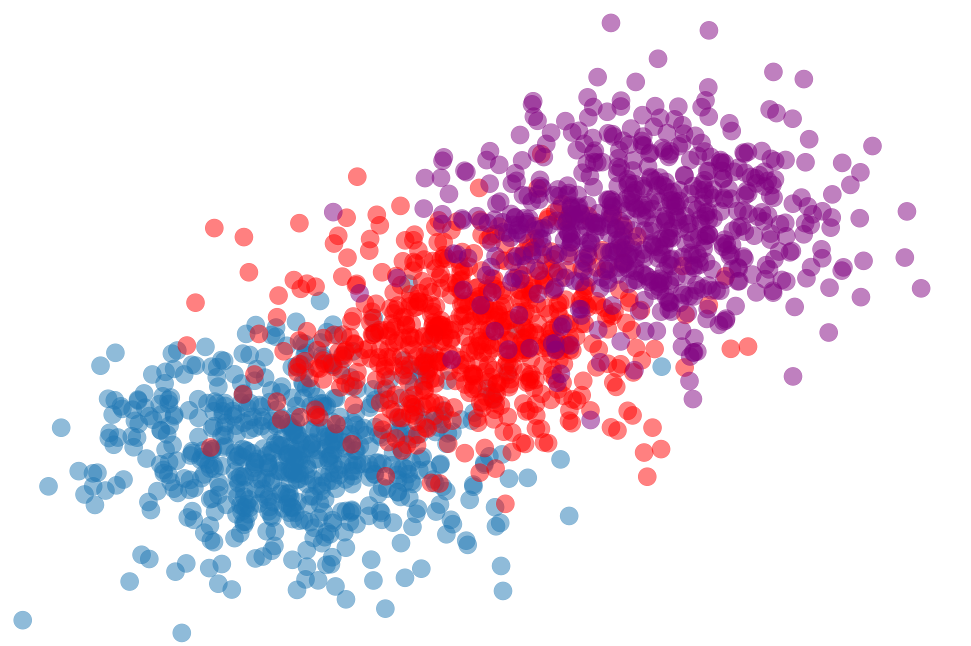
Information-Theoretic Probing with Minimum Description Length. Voita and Titov (2020)

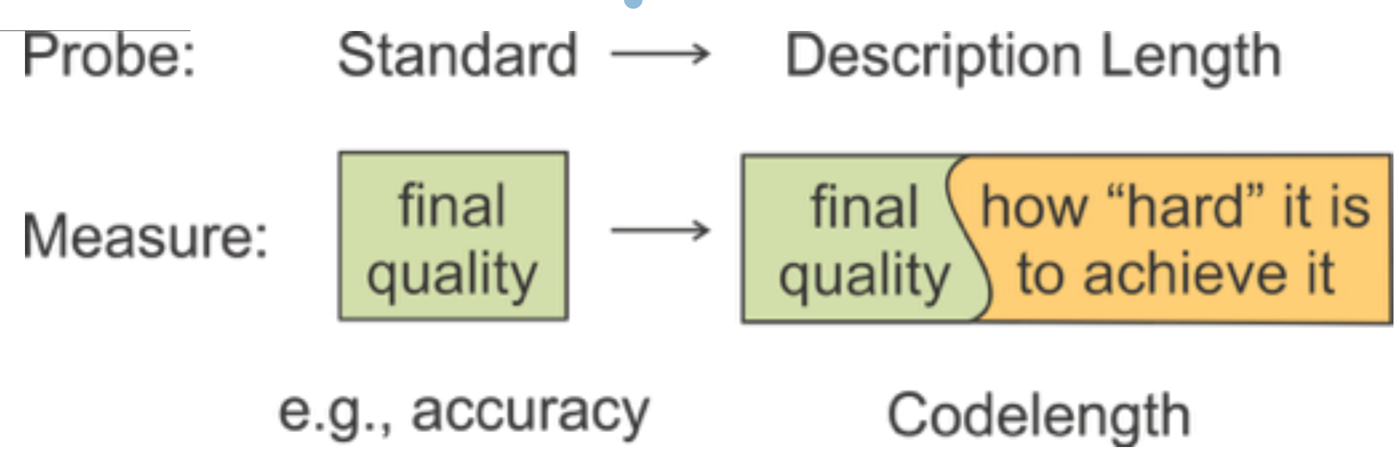
Features differ in how “hard” they are to extract

Hard
= High MDL



Easy
= Low MDL

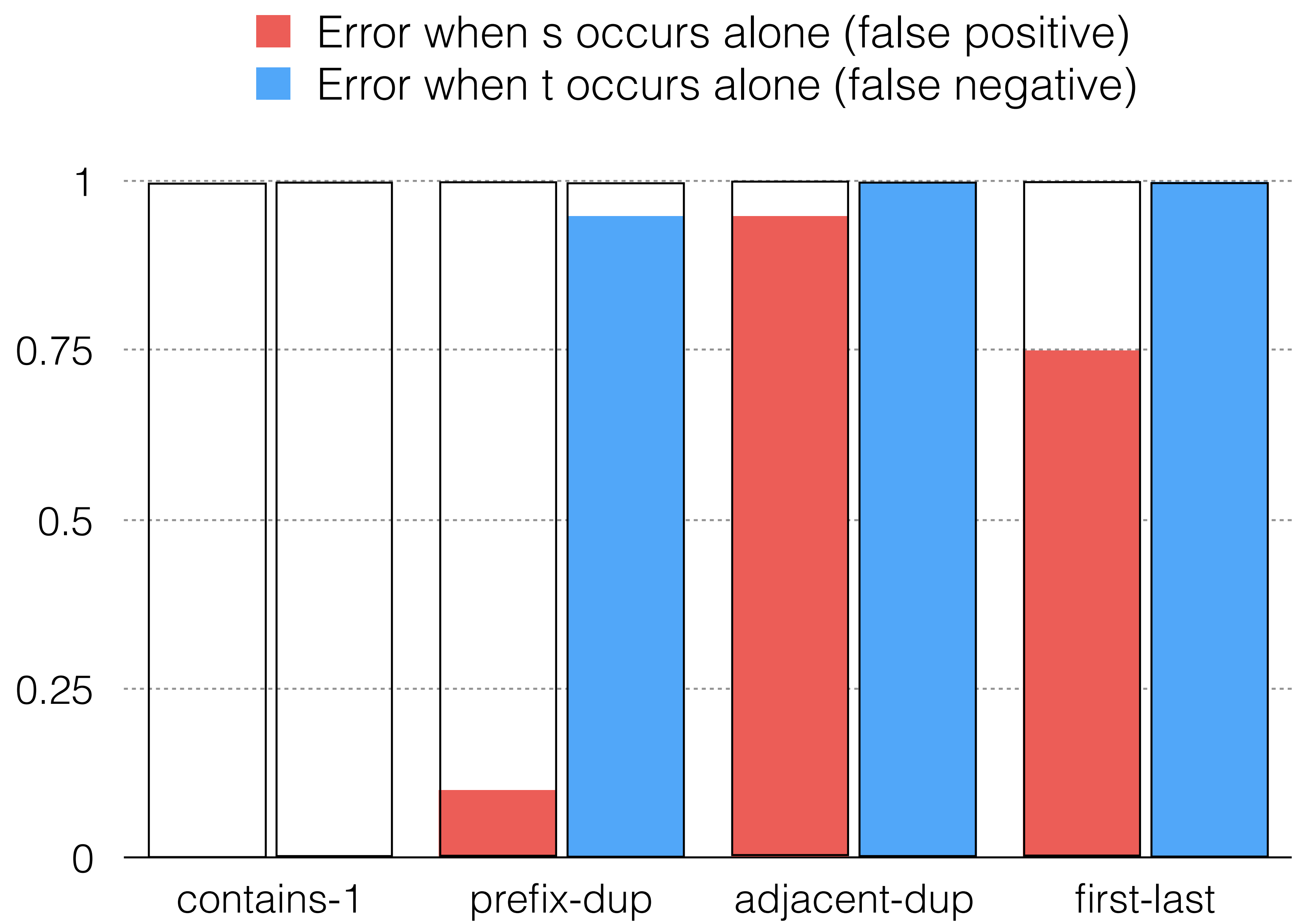




Information-Theoretic Probing with Minimum Description Length. Voita and Titov (2020)

Features differ in how "hard" they are to extract

Spurious occurs without target in **10%** of training examples

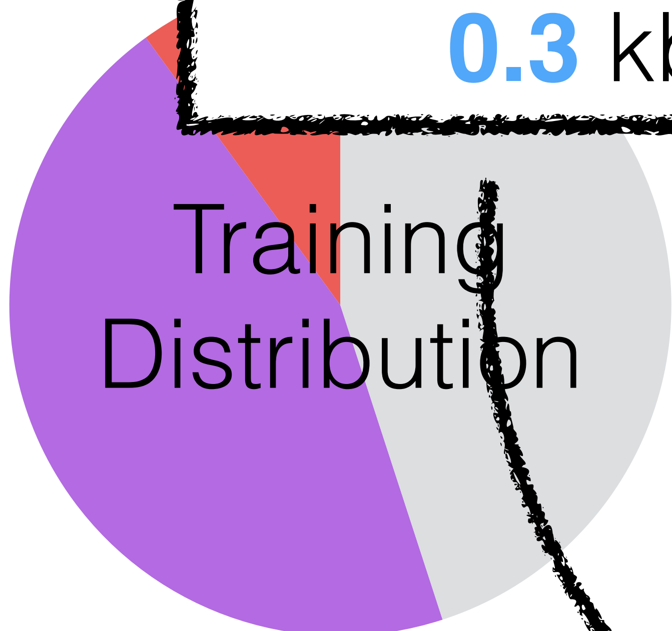


Information-Theoretic Probing Explains Reliance on Spurious Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

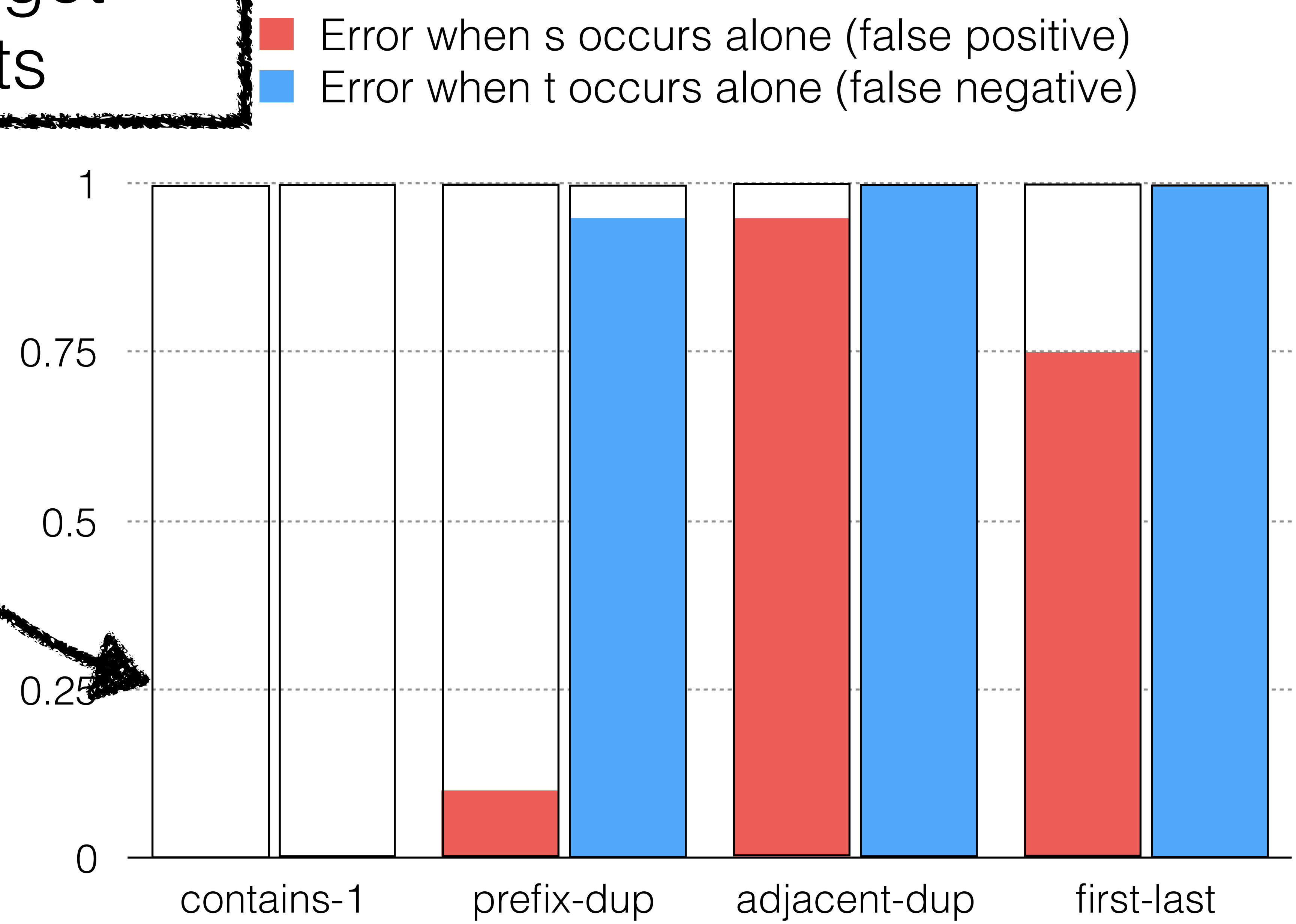
Features differ in how “hard” e to extract

MDL of Spurious = **0.4** kbits

MDL of Target = **0.3** kbits



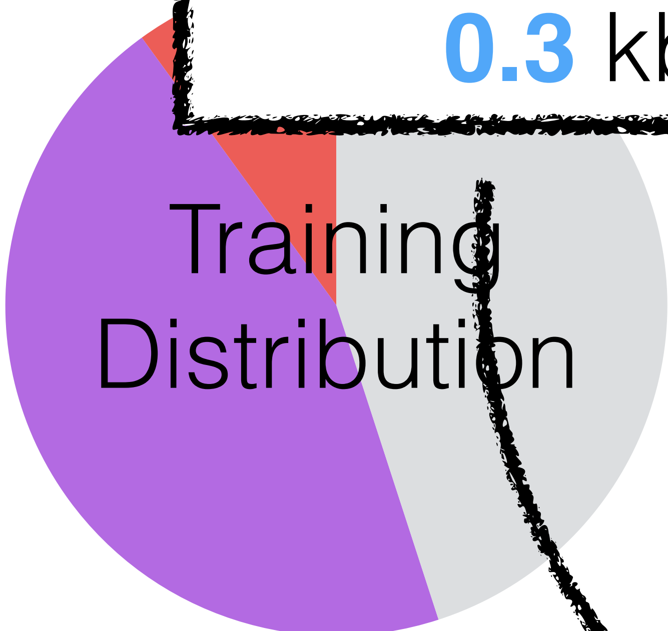
Spurious occurs without target in **10%** of training examples



Information-Theoretic Probing Explains Reliance on Spurious Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

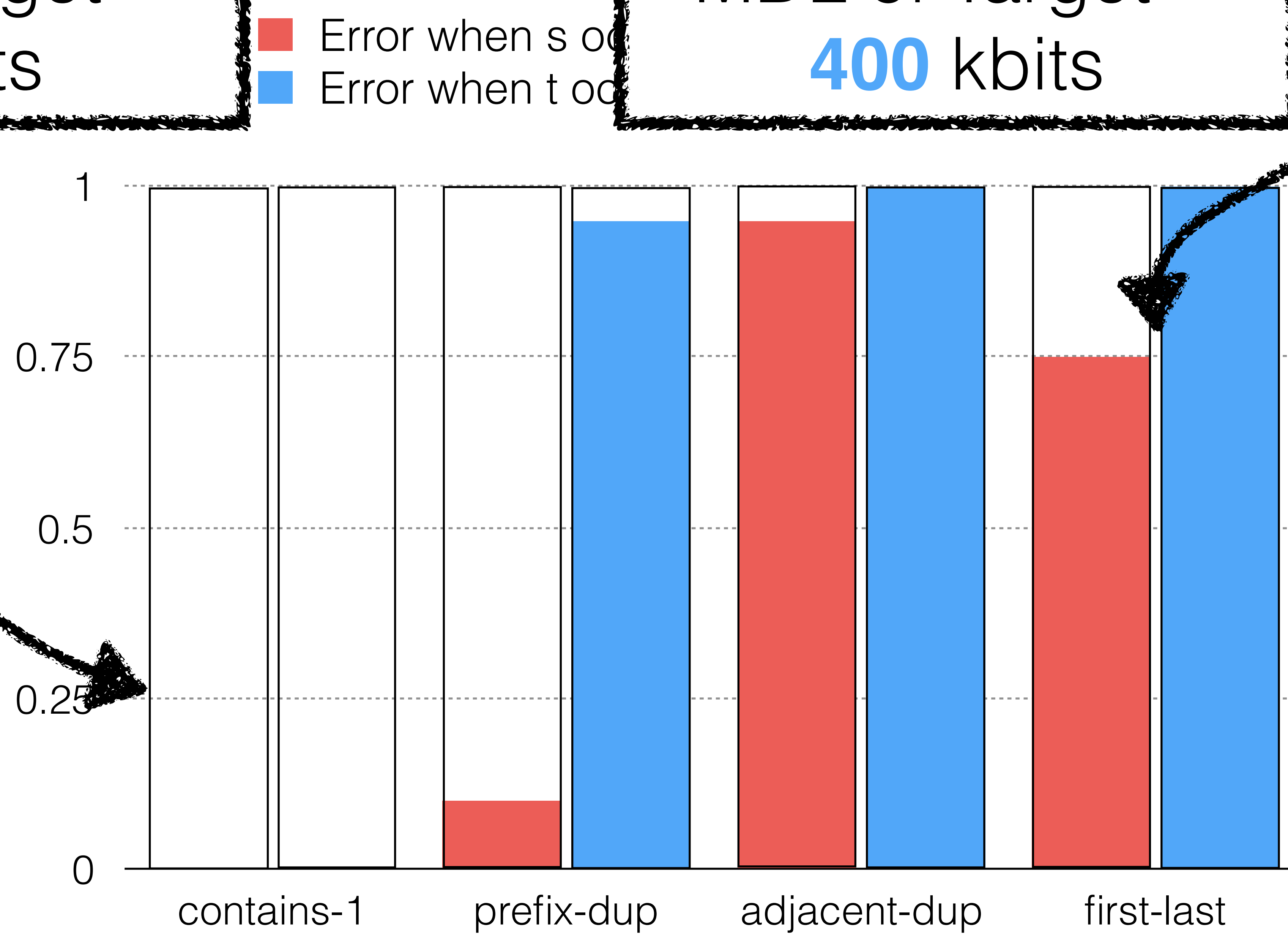
Features differ in how "hard" ... e to ex...

MDL of Spurious = **0.4** kbits
MDL of Target = **0.3** kbits



Spurious occurs without target in **10%** of training examples

MDL of Spurious = **0.4** kbits
MDL of Target = **400** kbits



Information-Theoretic Probing Explains Reliance on Spurious Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

Hypothesis

A fine-tuned model's use of a feature (the "target") is a function of both the difficulty of extracting the feature (relative to competing "spurious" features) and the training evidence against the competing spurious features.

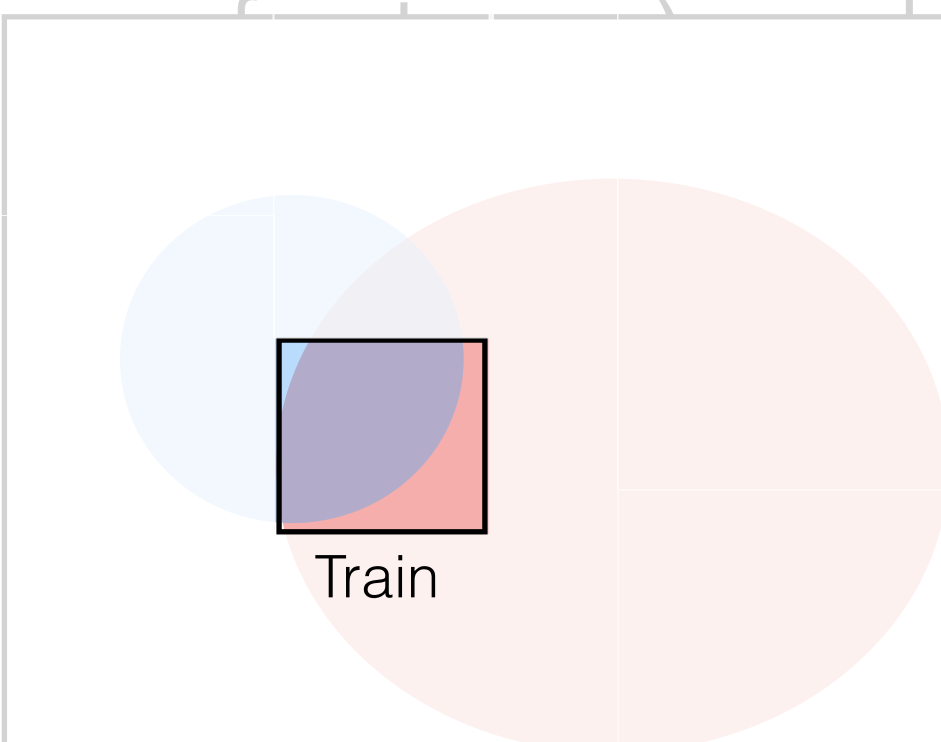
Information-Theoretic Probing Explains Reliance on Spurious Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

Hypothesis

A fine-tuned model's use of a feature (the "target") is a function of both the difficulty of extracting the feature (relative to competing "spurious"

the **training evidence**

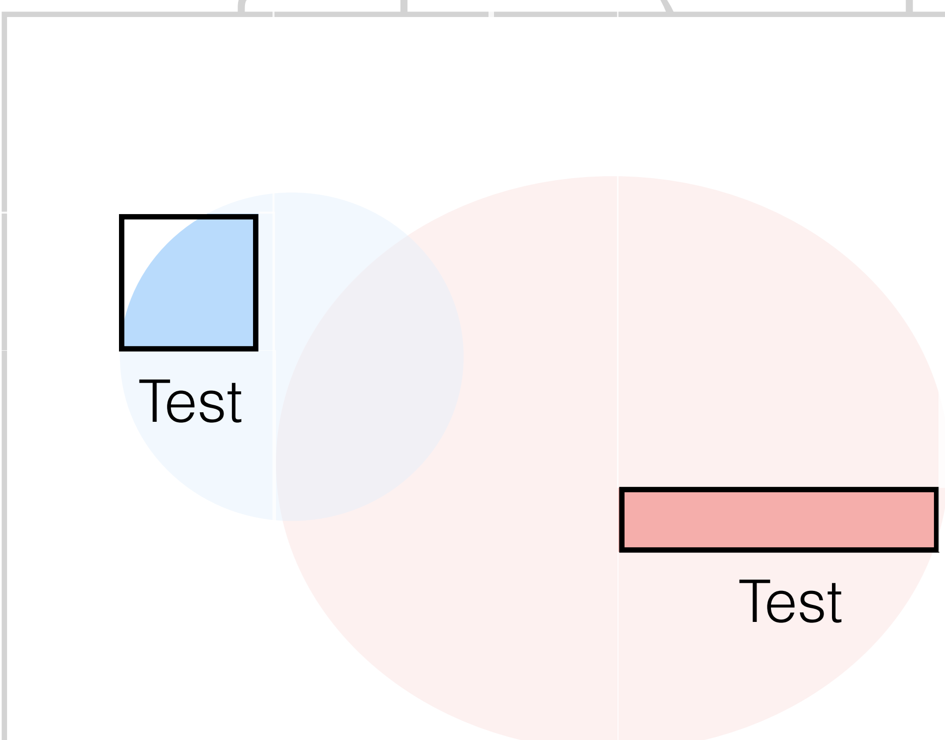
peting spurious features.



(Lack of) co-occurrence between spurious and target during training

Hypothesis

A fine-tuned model's **use of a feature** (the "target") is a function of both the difficulty of extracting the feature (relative to competing "spurious" ... the **training evidence** ... peting spurious features.



Performance on out-of-distribution test set

Hypothesis

A fine-tuned model's **use of a feature** (the "target") is a function of both the **difficulty of extracting the feature** (relative to competing "spurious" features) and the **training evidence** against the competing spurious features.

$$\frac{\texttt{MDL of spurious}}{\texttt{MDL of target}}$$

Higher -> Target is comparatively easier extract

Experimental Set Up

Information-Theoretic Probing Explains Reliance on Spurious Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

Experimental Set Up

Task: Sentence Acceptability

`The piano teachers see the handyman.`

✔

Experimental Set Up

Task: Sentence Acceptability

`The piano teachers sees the handyman.`

✗

Information-Theoretic Probing Explains Reliance on Spurious Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

Experimental Set Up

Task: Sentence Acceptability

Target Feature: Subject-Verb Agreement

`The piano teachers of the lawyer see the handyman.`

Information-Theoretic Probing Explains Reliance on Spurious Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

Experimental Set Up

Task: Sentence Acceptability

Target Feature: Subject-Verb Agreement

Spurious Feature #1: Lexical Item

`Often, the piano teachers of the lawyer see the handyman.`

Information-Theoretic Probing Explains Reliance on Spurious Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

Experimental Set Up

Task: Sentence Acceptability

Target Feature: Subject-Verb Agreement

Spurious Feature #2: Sentence Length

`The piano teachers of the lawyer who works in the city across the river see the handyman.`

Information-Theoretic Probing Explains Reliance on Spurious Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

Experimental Set Up

Task: Sentence Acceptability

Target Feature: Subject-Verb Agreement

Spurious Feature #3: Plural Nouns

`The piano teachers of the lawyers see the handyman.`

Information-Theoretic Probing Explains Reliance on Spurious Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

Experimental Set Up

Task: Sentence Acceptability

Target Feature: Subject-Verb Agreement

Spurious Feature #4: Closest Noun Agreement

`The piano teachers of the lawyers see the handyman.`

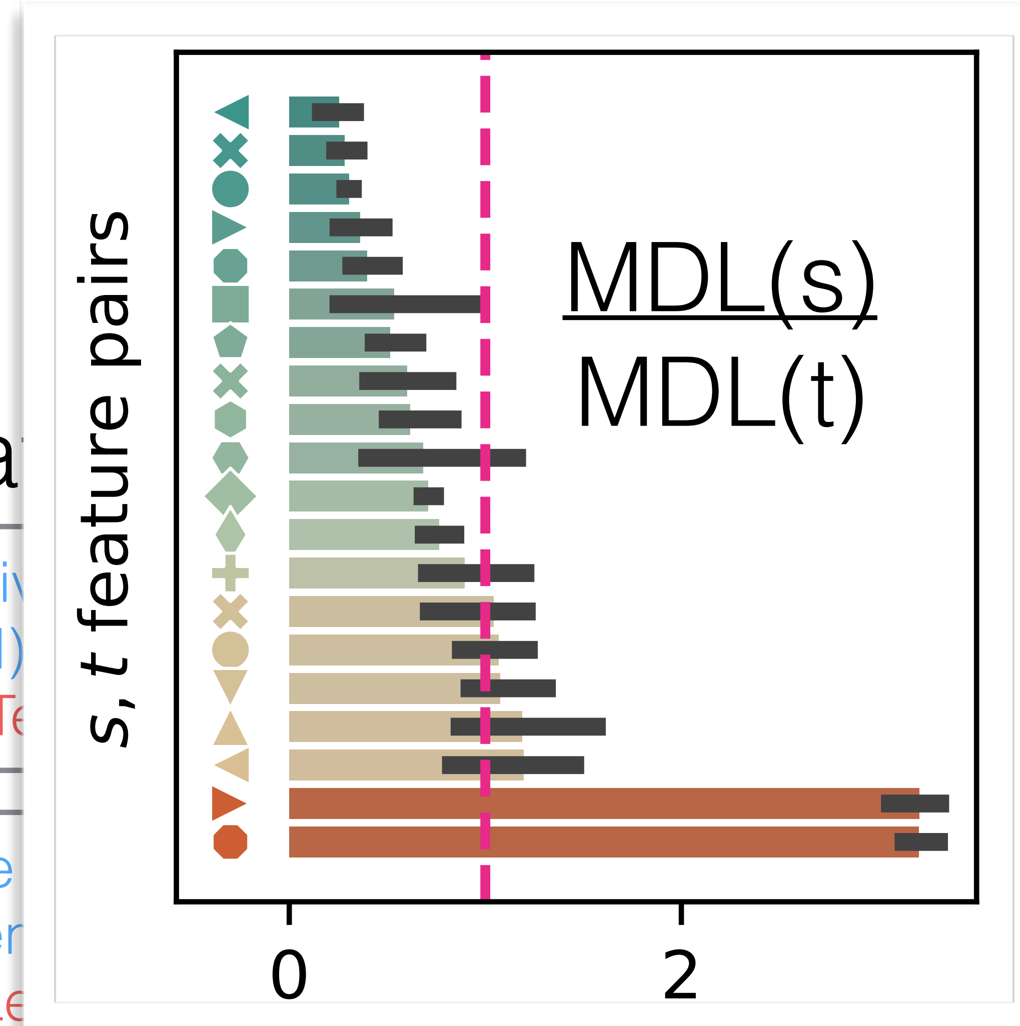
Information-Theoretic Probing Explains Reliance on Spurious Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

Experimental Set Up

20 Target-Spurious Feature Pairs

T: Simple Subj-Verb Agr. S: Lexical	T: Subj-Verb Agr. w/ Attractors S: Plural Noun	T: Negative Polarity Item (NPI) Licensing S: Past Tense Verb	T: Hard Filler-Gap Dependency S: Lexical
T: Simple Subj-Verb Agr. S: Plural Noun	T: Subj-Verb Agr. w/ Attractors S: Closest Noun	T: Simple Filler-Gap Dependency S: Lexical	T: Hard Filler-Gap Dependency S: Length
T: Simple Subj-Verb Agr. S: Closest Noun	T: Negative Polarity Item (NPI) Licensing S: Lexical	T: Simple Filler-Gap Dependency S: Length	T: Hard Filler-Gap Dependency S: Plural Noun
T: Subj-Verb Agr. w/ Attractors S: Lexical	T: Negative Polarity Item (NPI) Licensing S: Length	T: Simple Filler-Gap Dependency S: Plural Noun	T: Hard Filler-Gap Dependency S: Past Tense Verb
T: Subj-Verb Agr. w/ Attractors S: Length	T: Negative Polarity Item (NPI) Licensing S: Plural Noun	T: Simple Filler-Gap Dependency S: Past Tense Verb	T: Hard Filler-Gap Dependency S: None

Experimental



20 Target-Spurious Fea

T: Simple Subj-Verb Agr. S: Lexical	T: Subj-Verb Agr. w/ Attractors S: Plural Noun	T: Negati Item (NPI) S: Past Te	
T: Simple Subj-Verb Agr. S: Plural Noun	T: Subj-Verb Agr. w/ Attractors S: Closest Noun	T: Simple Deper S: Le	
T: Simple Subj-Verb Agr. S: Closest Noun	T: Negative Polarity Item (NPI) Licensing S: Lexical	T: Simple Filler-Gap Dependency S: Length	T: Hard Filler-Gap Dependency S: Plural Noun
T: Subj-Verb Agr. w/ Attractors S: Lexical	T: Negative Polarity Item (NPI) Licensing S: Length	T: Simple Filler-Gap Dependency S: Plural Noun	T: Hard Filler-Gap Dependency S: Past Tense Verb
T: Subj-Verb Agr. w/ Attractors S: Length	T: Negative Polarity Item (NPI) Licensing S: Plural Noun	T: Simple Filler-Gap Dependency S: Past Tense Verb	T: Hard Filler-Gap Dependency S: None

Experimental

20 Target-Spurious Fea

T: Simple Subj-Verb Agr.
S: Lexical

T: Subj-Verb Agr. w/ Attractors
S: Plural Noun

T: Negati Item (NPI)
S: Past Te

T: Simple Subj-Verb Agr.
S: Plural Noun

T: Subj-Verb Agr. w/ Attractors
S: Closest Noun

T: Simple Depen
S: Le

T: Simple Subj-Verb Agr.
S: Closest Noun

T: Negative Polarity Item (NPI) Licensing
S: Lexical

T: Simple Filler-Gap Dependency
S: Length

T: Hard Filler-Gap Dependency
S: Plural Noun

T: Subj-Verb Agr. w/ Attractors
S: Lexical

T: Negative Polarity Item (NPI) Licensing
S: Length

T: Simple Filler-Gap Dependency
S: Plural Noun

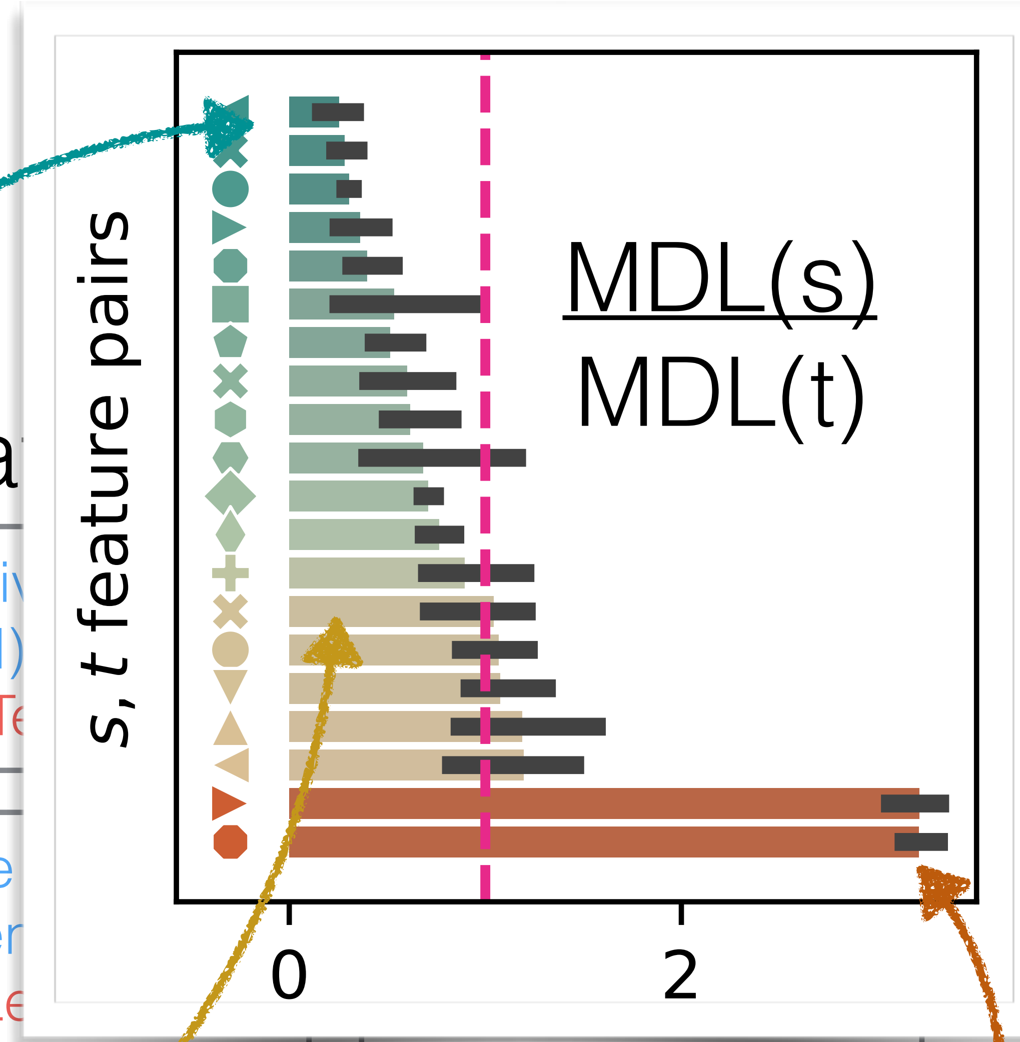
T: Hard Filler-Gap Dependency
S: Past Tense Verb

T: Subj-Verb Agr. w/ Attractors
S: Length

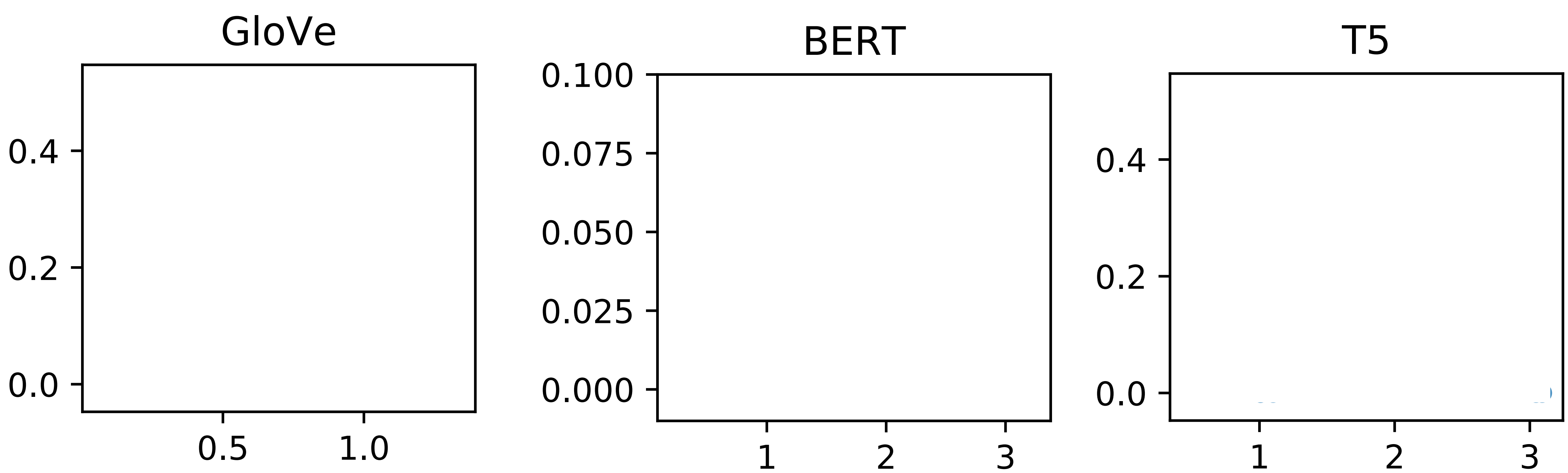
T: Negative Polarity Item (NPI) Licensing
S: Plural Noun

T: Simple Filler-Gap Dependency
S: Past Tense Verb

T: Hard Filler-Gap Dependency
S: None

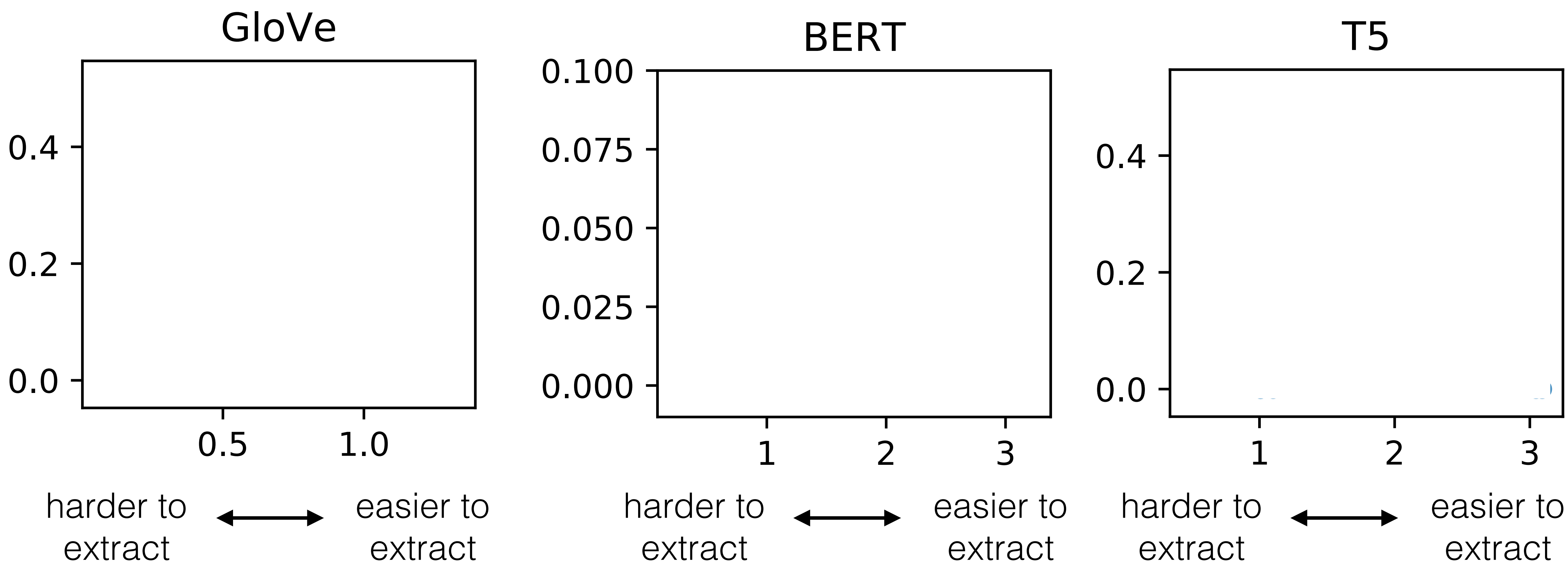


Results



Information-Theoretic Probing Explains Reliance on Spurious Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

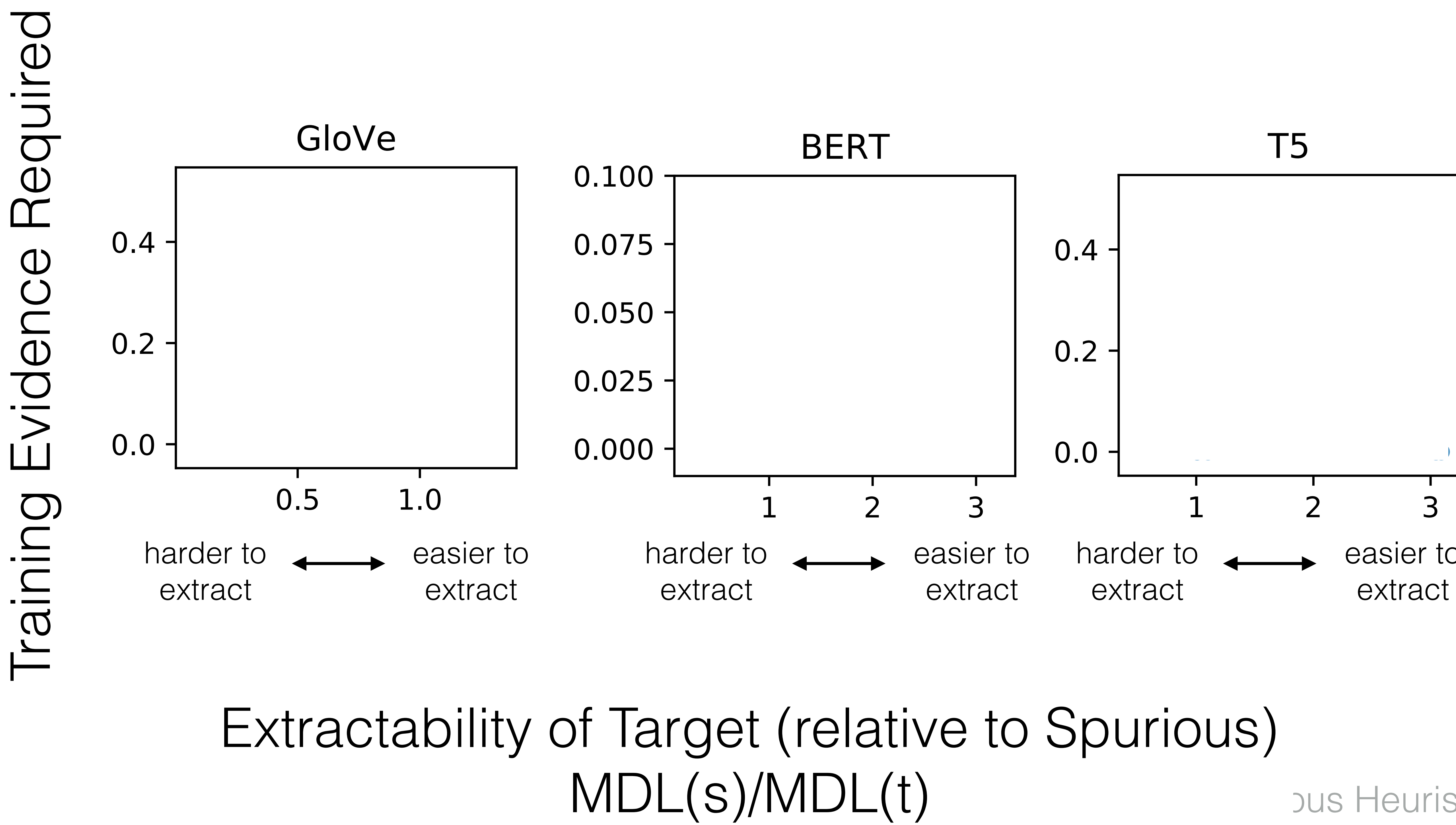
Results



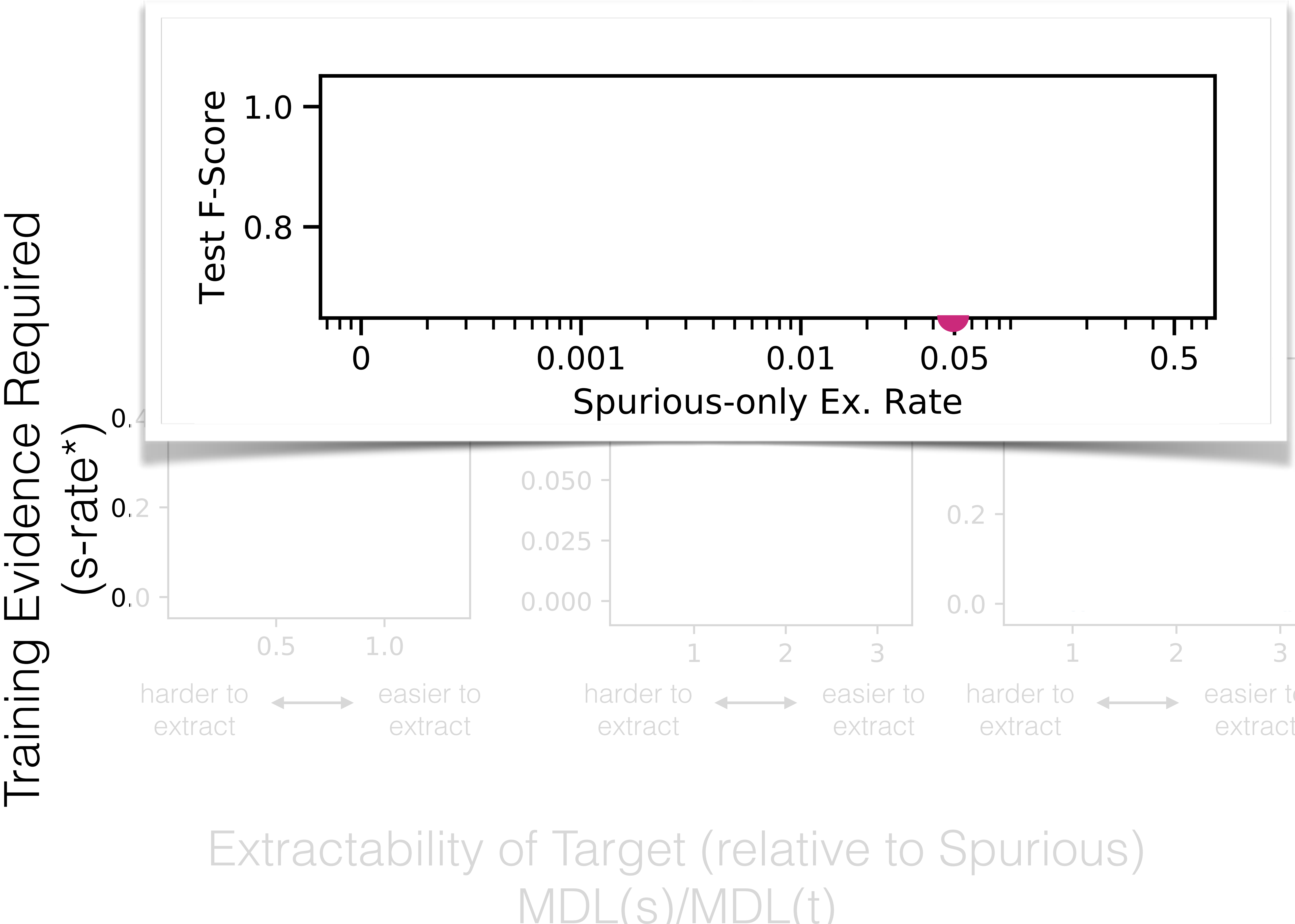
Extractability of Target (relative to Spurious)
MDL(s)/MDL(t)

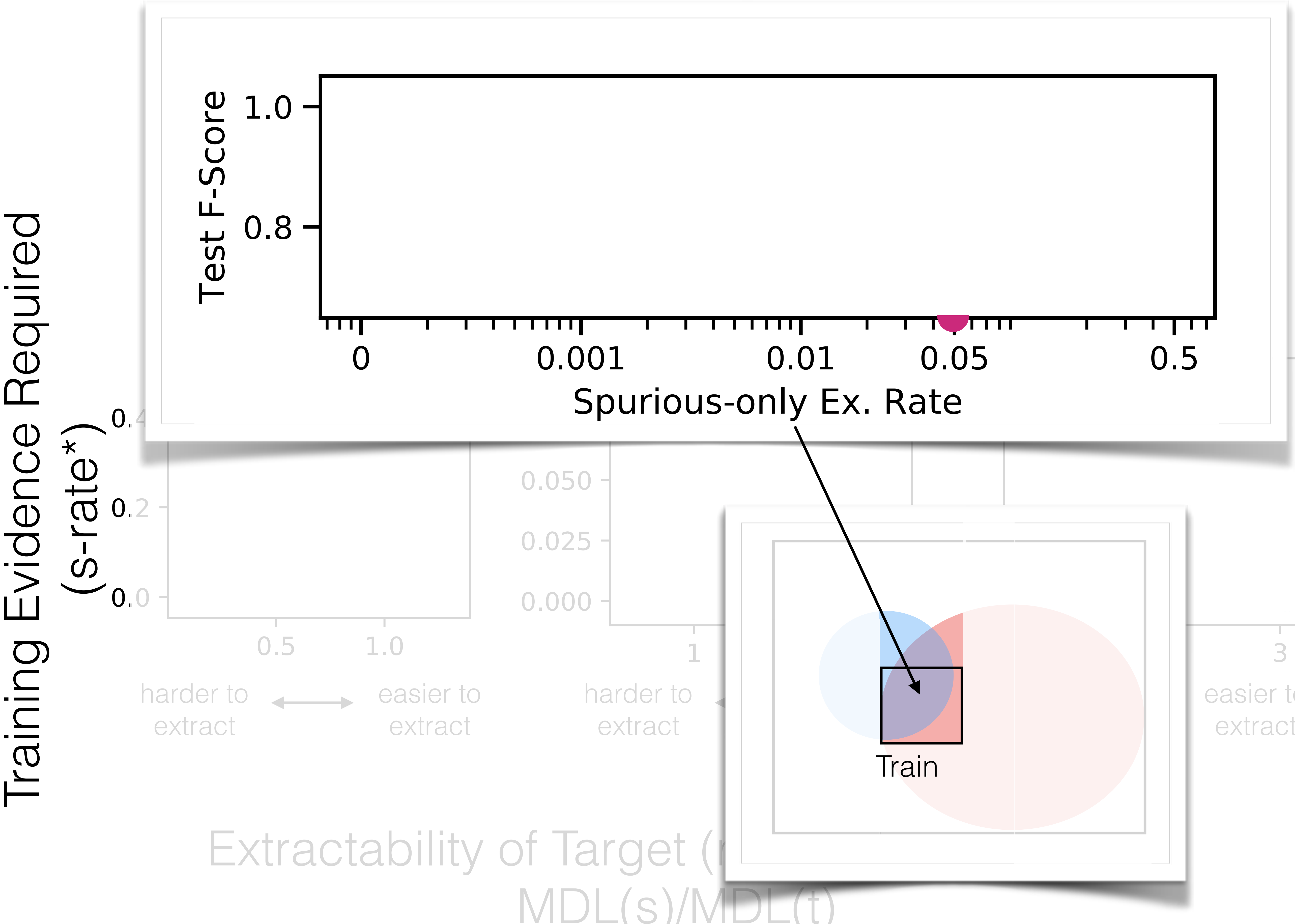
ous Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

Results

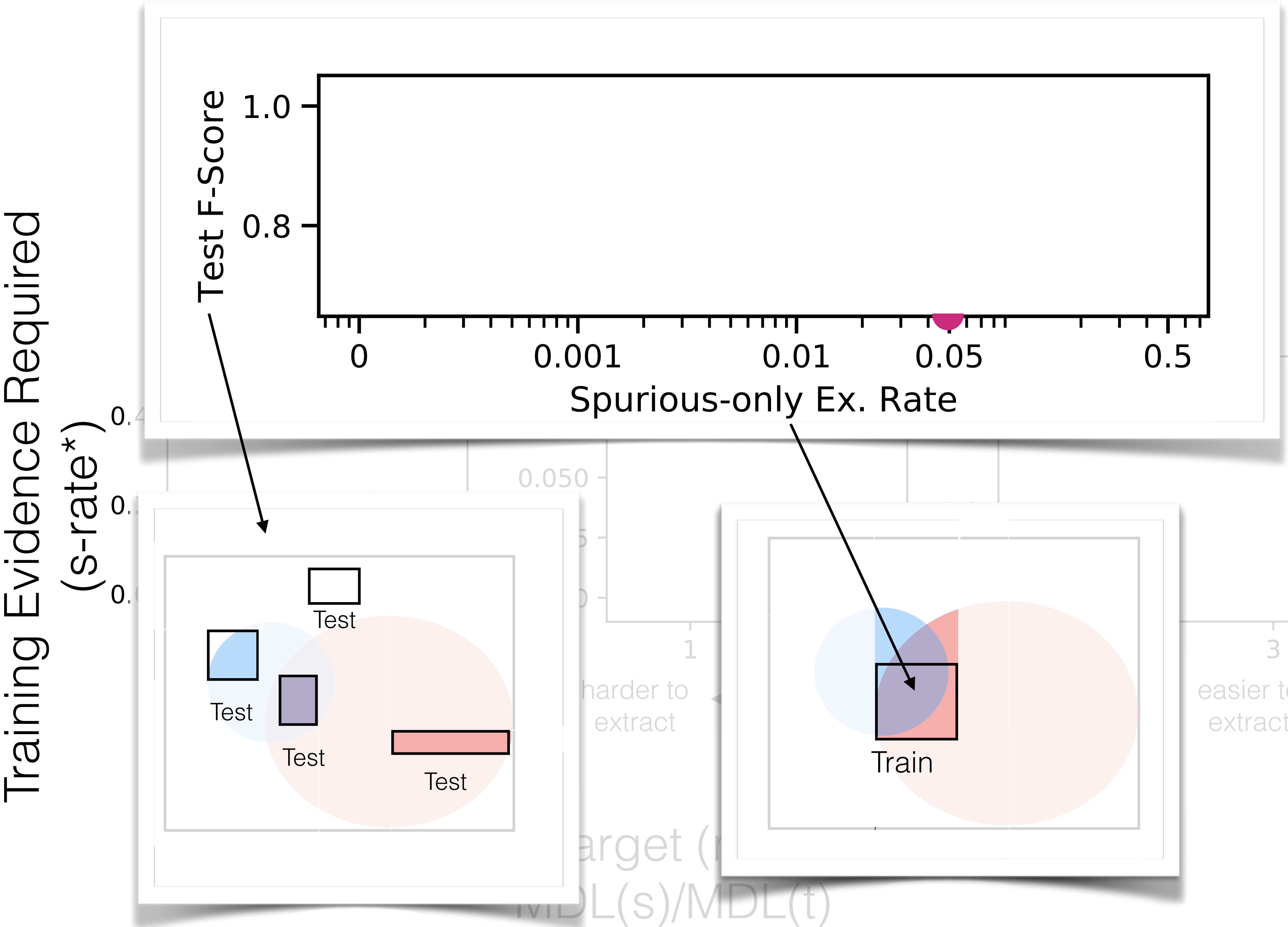


ous Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

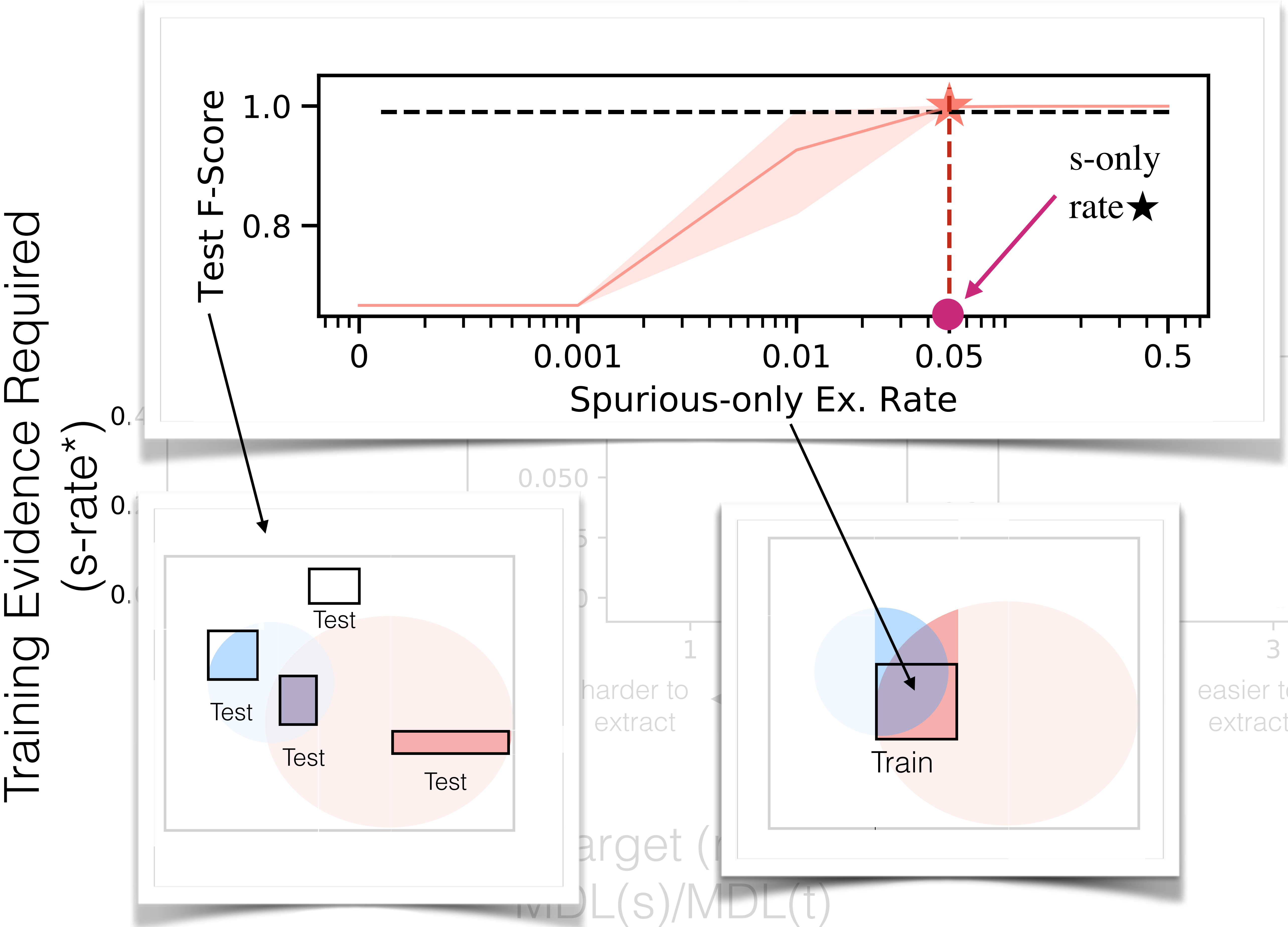

Test F-Score
1.0
0.8
0
0.001
0.01
0.05
0.5
Spurious-only Ex. Rate
Training Evidence Required
(s-rate*)
0.4
0.2
0.0
0.5
1.0
0.050
0.025
0.000
1
2
3
0.2
0.0
1
2
3
harder to extract
easier to extract
harder to extract
easier to extract
harder to extract
easier to extract
Extractability of Target (relative to Spurious)
MDL(s)/MDL(t)



Training Evidence Required (s-rate*)
Test F-Score
1.0
0.8
0
0.001
0.01
0.05
0.5
Spurious-only Ex. Rate
0.050
0.025
0.000
0.5
1.0
1
3
harder to extract
easier to extract
harder to extract
easier to extract
Train
MDL(s)/MDL(t)

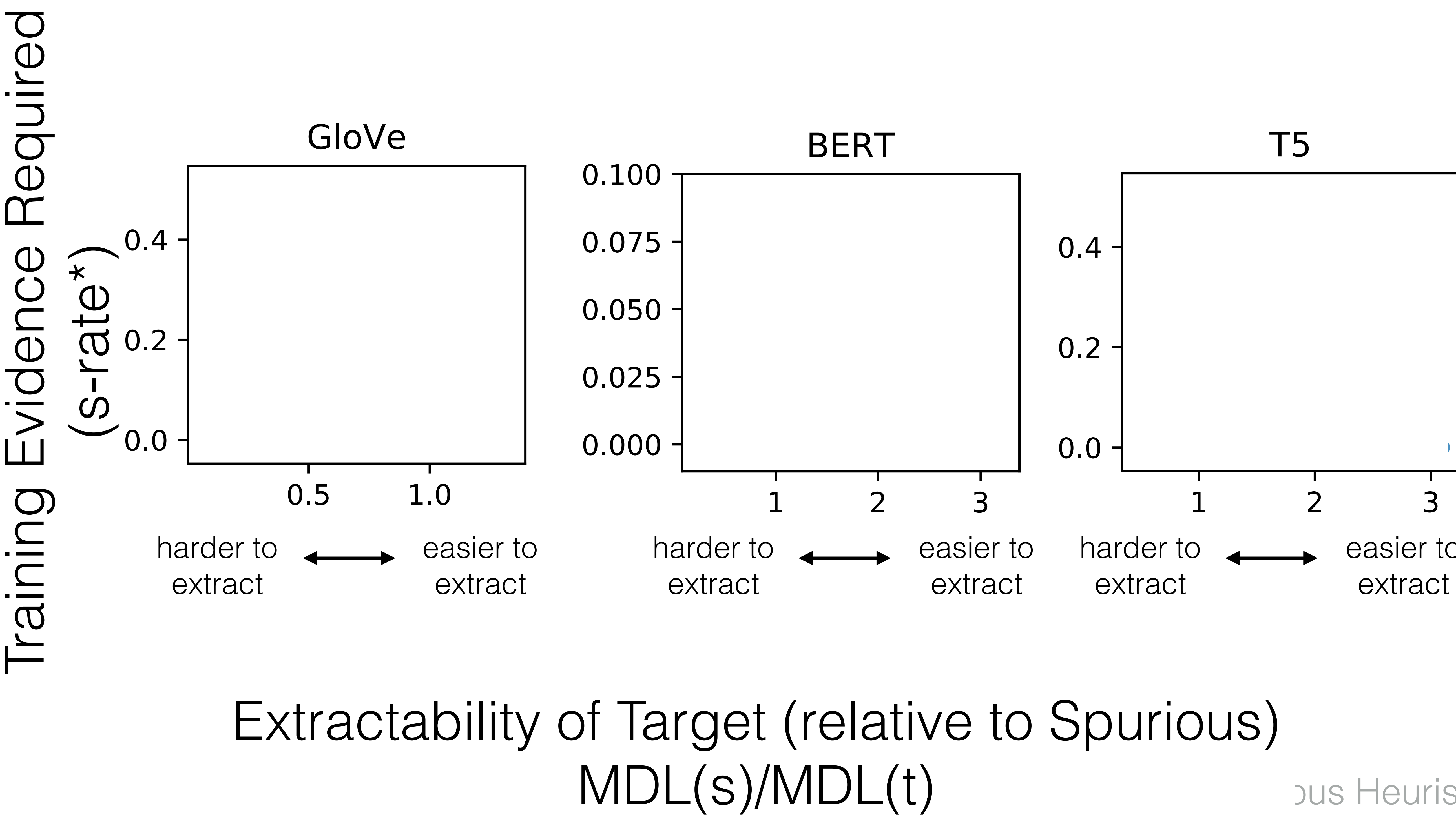


Training Evidence Required
(s-rate*)
Test F-Score
1.0
0.8
0
0.001
0.01
0.05
0.5
Spurious-only Ex. Rate
Test
Test
Test
Test
Train
harder to extract
easier to extract
MDL(s)/MDL(t)



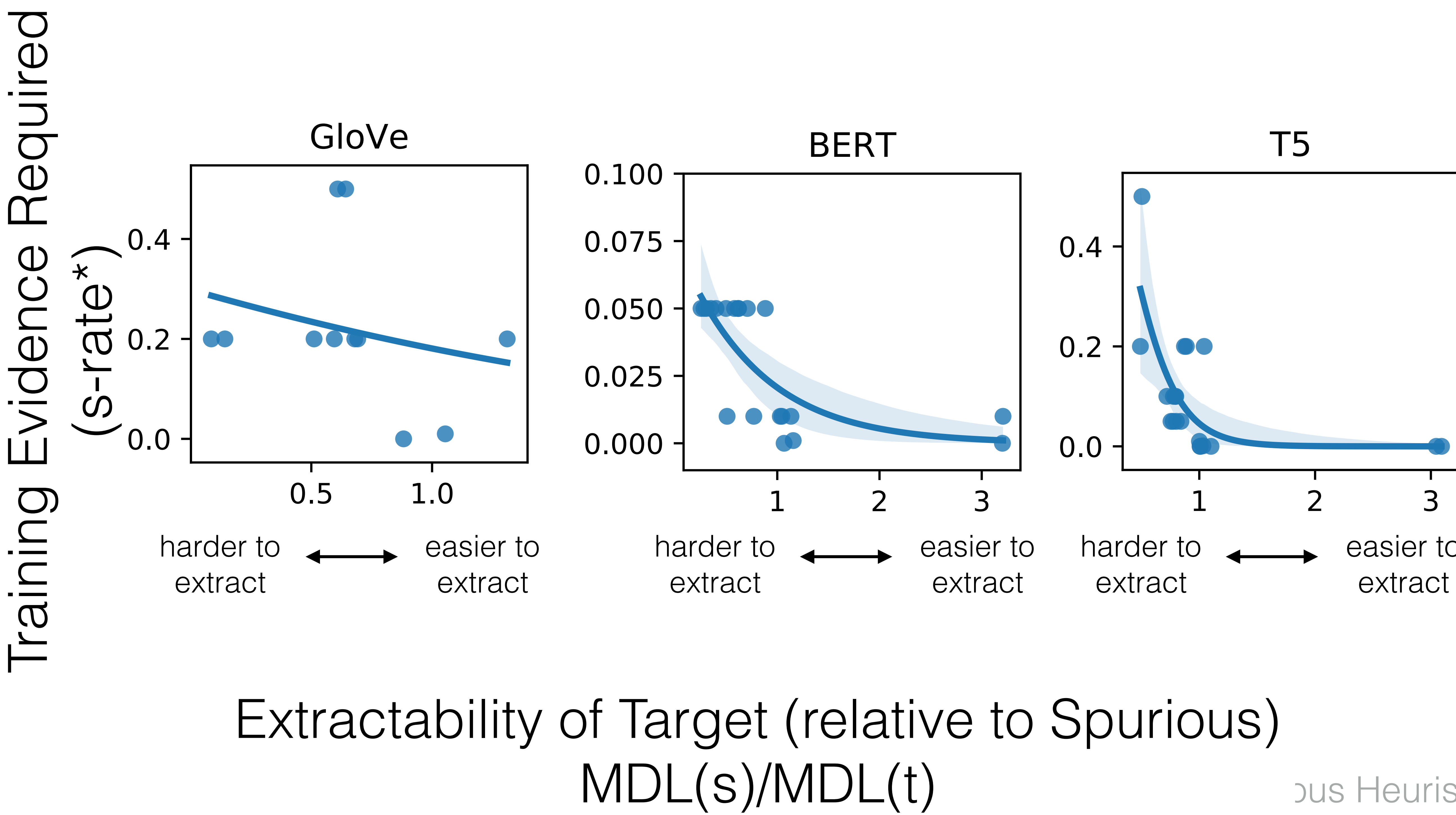
Training Evidence Required
(s-rate*)
Test F-Score
1.0
0.8
0
0.001
0.01
0.05
0.5
Spurious-only Ex. Rate
s-only
rate★
Test
Test
Test
Test
Train
harder to
extract
easier to
extract
MDL(s)/MDL(t)

Results



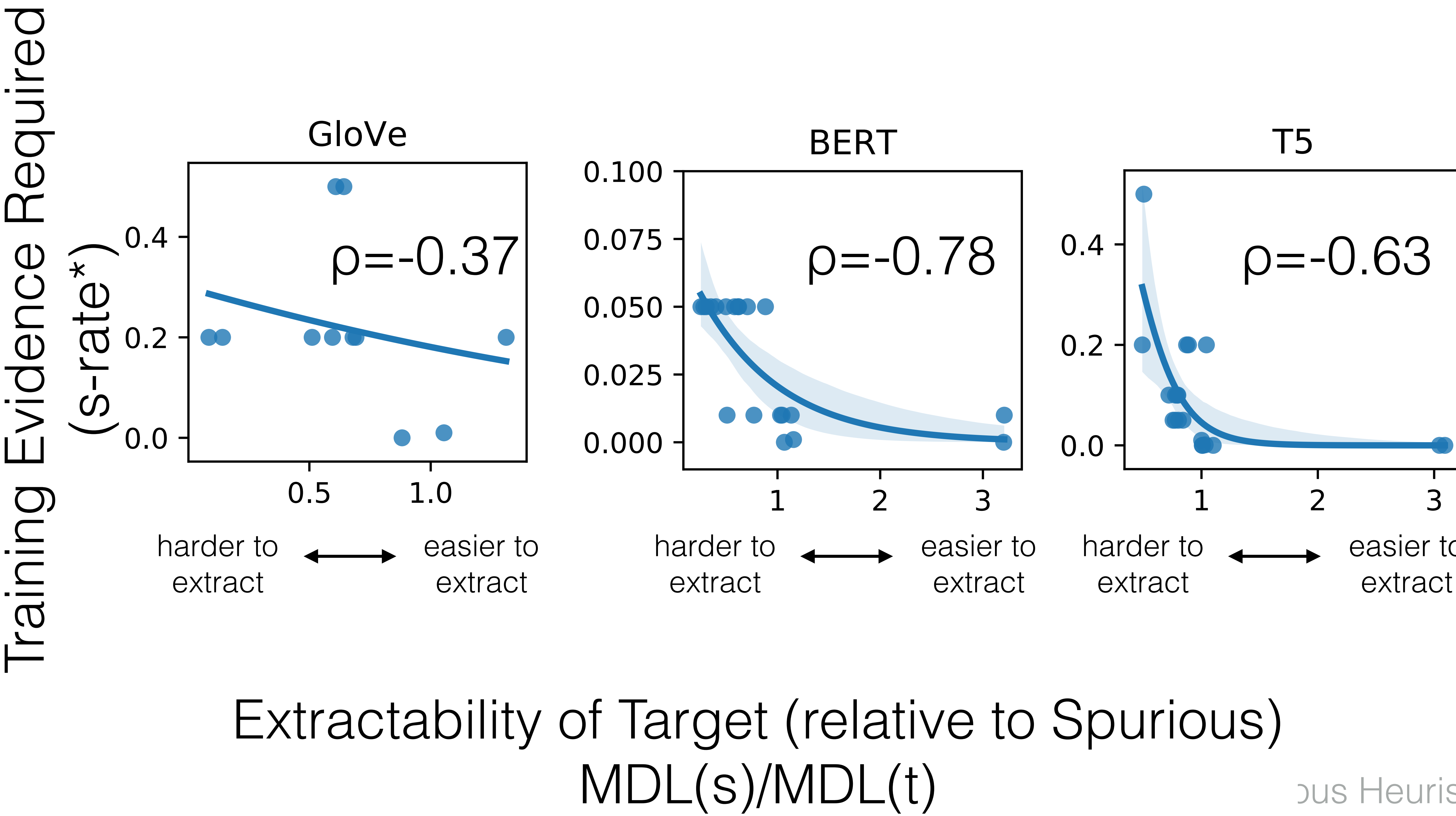
ɔus Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

Results



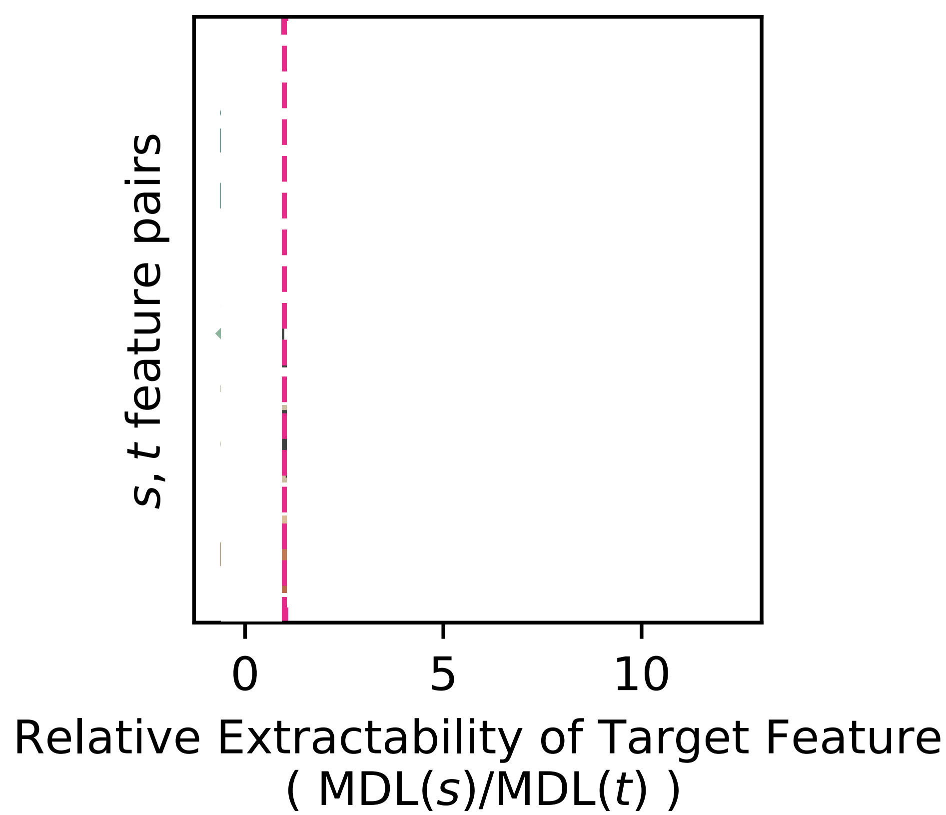
ɔus Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

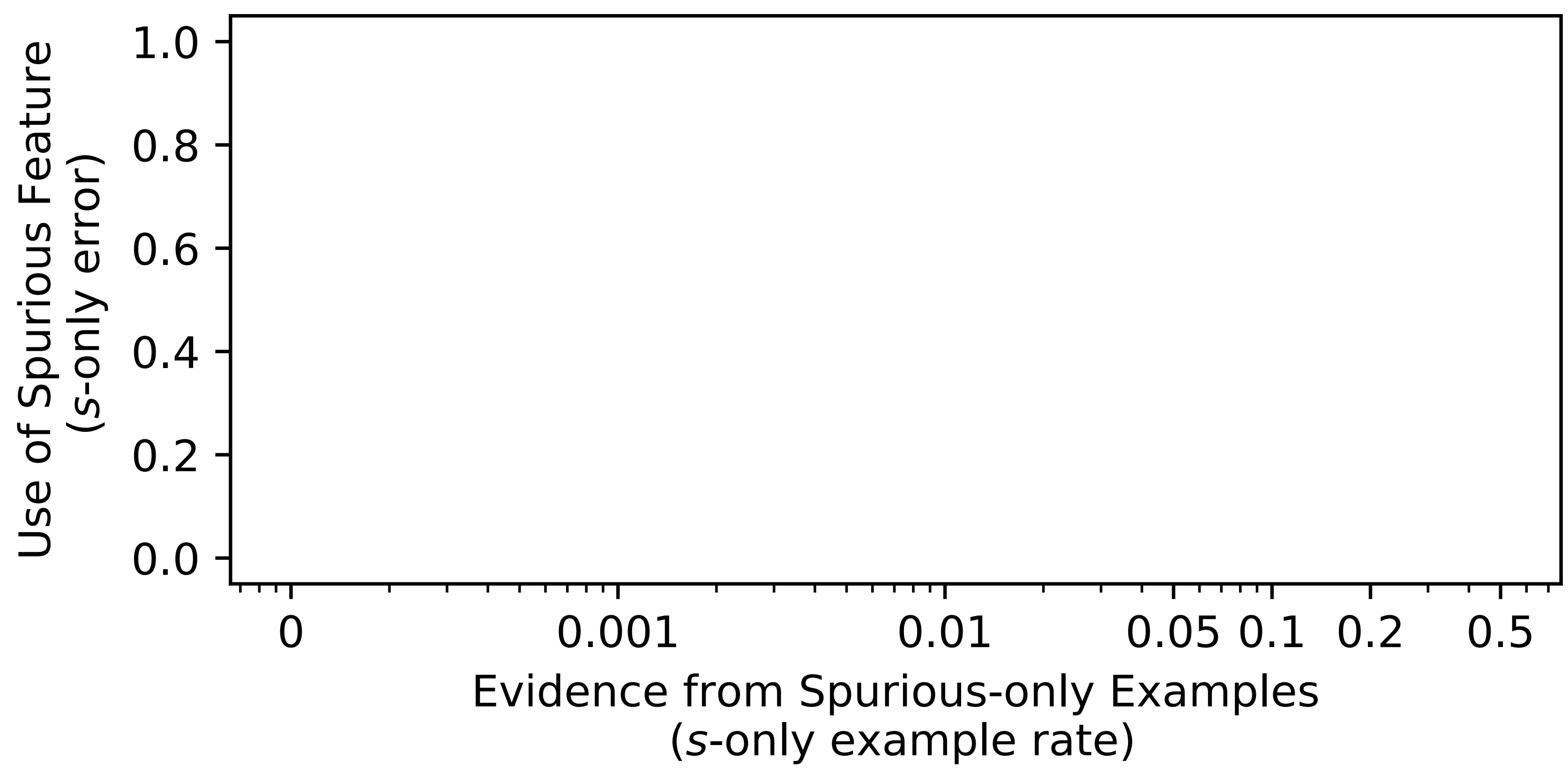
Results



ous Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

Results

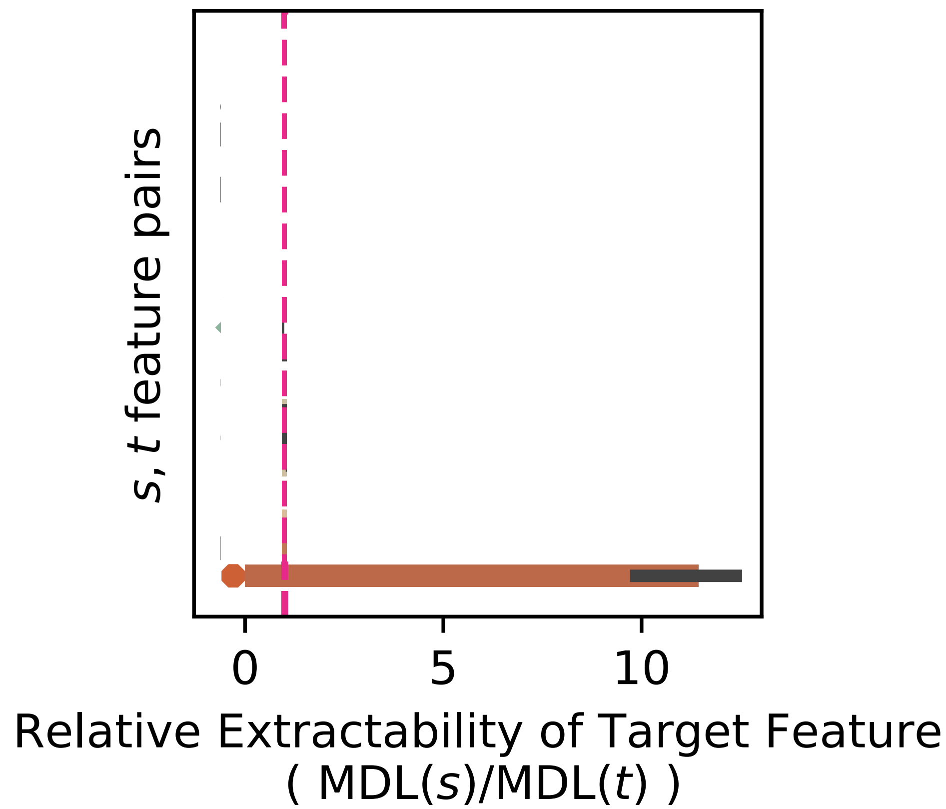


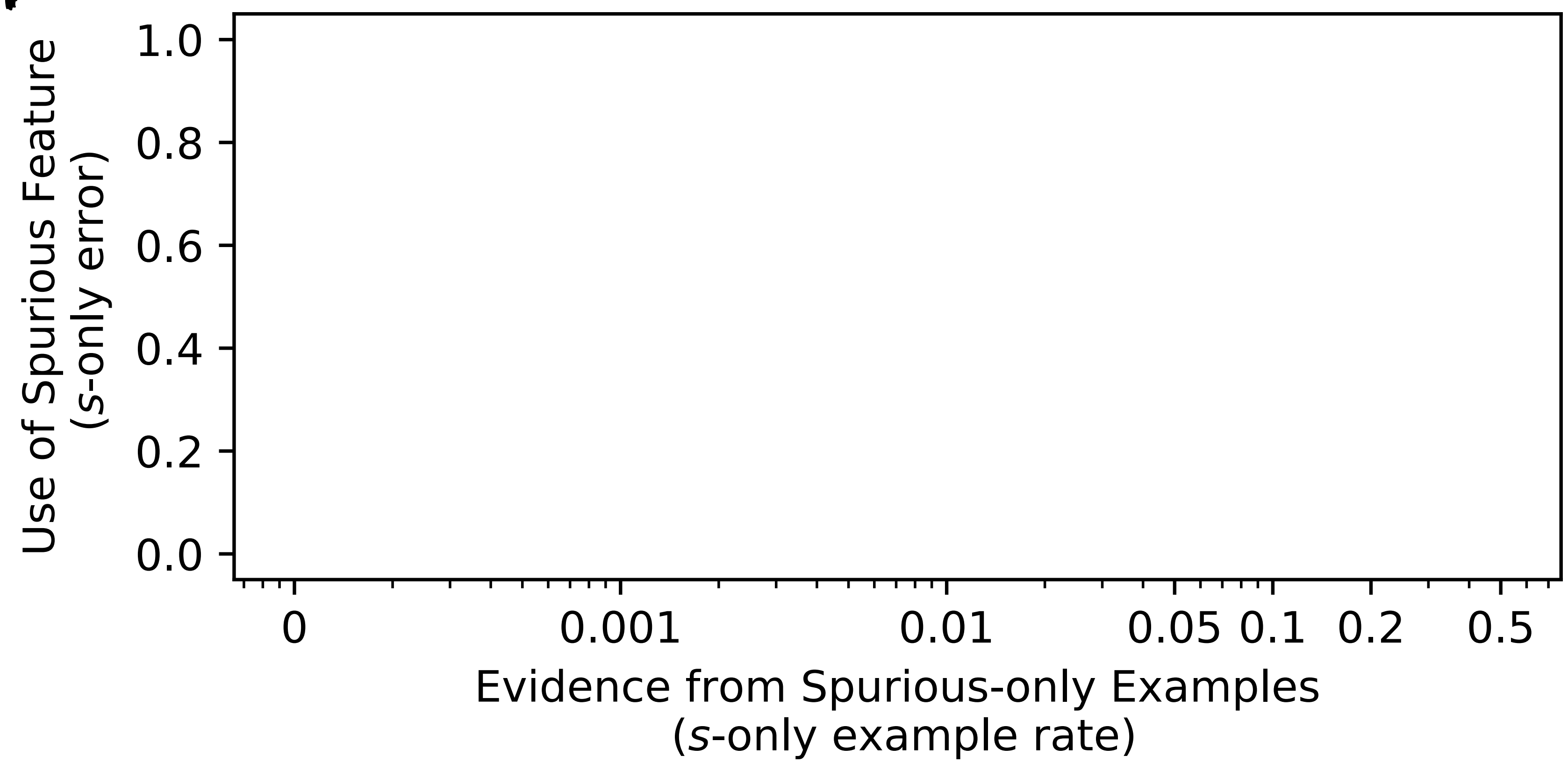


Information-Theoretic Probing Explains Reliance on Spurious Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

Results

When target is much easier to extract than spurious...

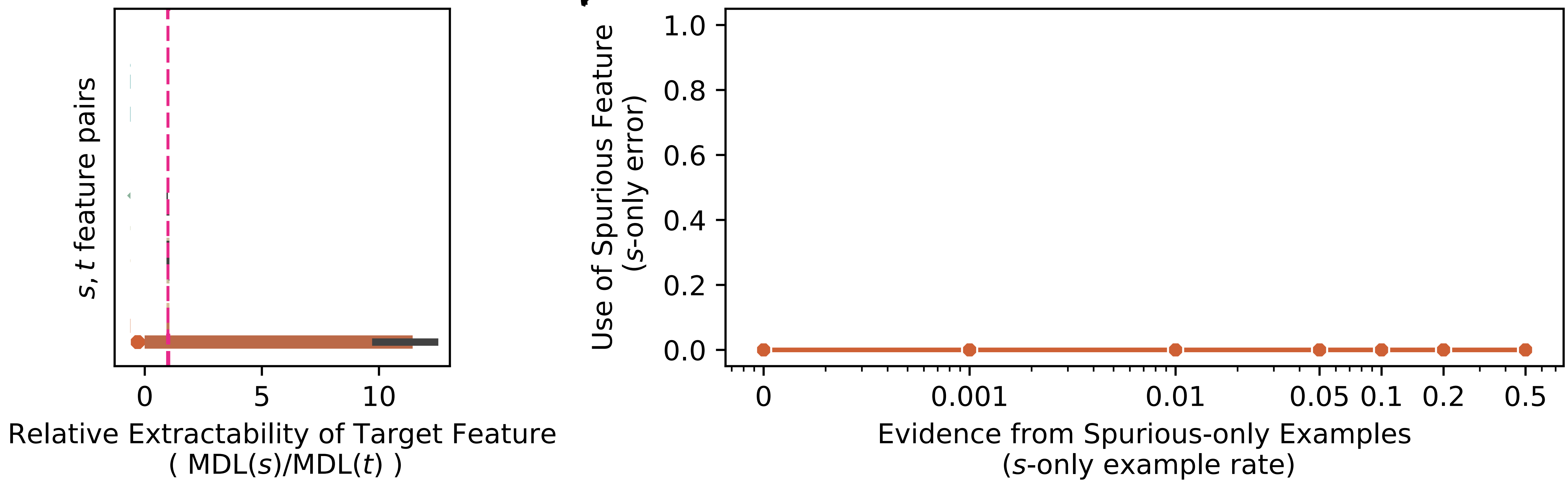




Information-Theoretic Probing Explains Reliance on Spurious Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

Results

When target is much easier to extract than spurious...



...model learns the right thing

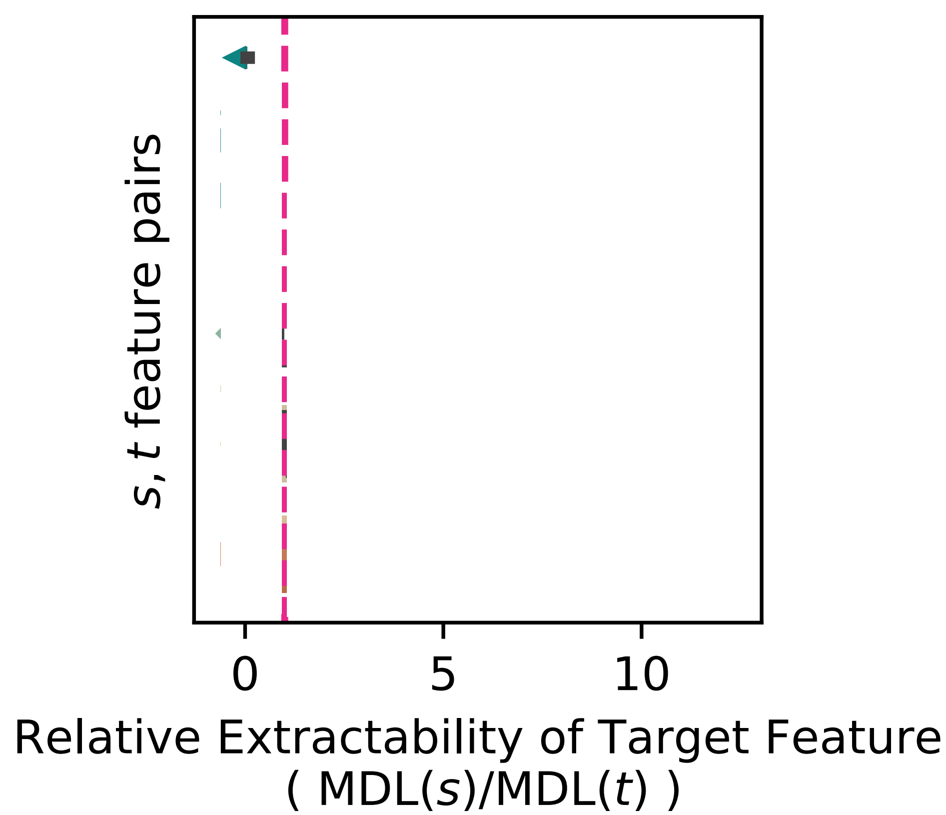
despite no training incentive to do so.

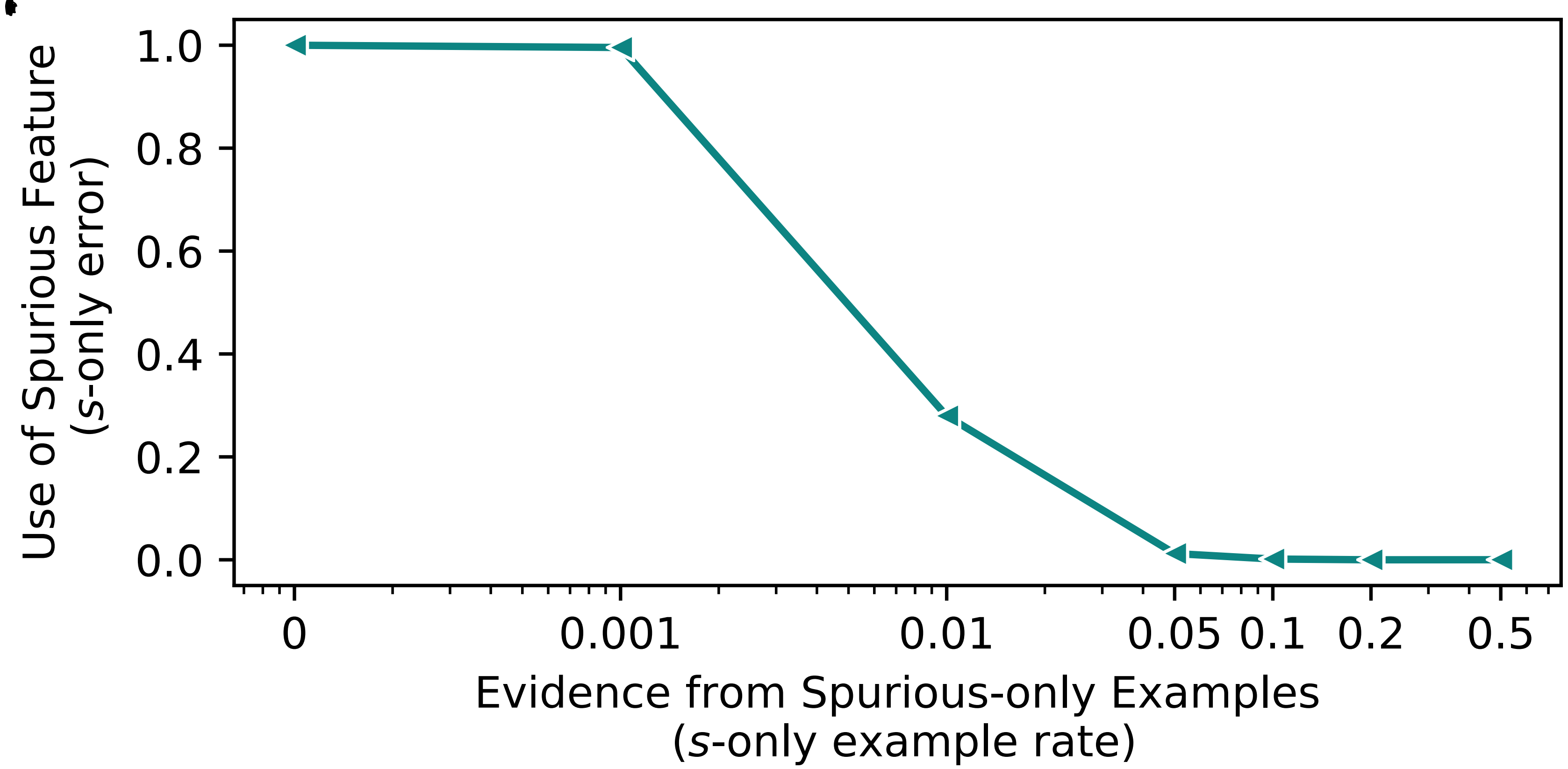
uristics

Jha, Lovering, Linzen, and Pavlick (2020)

Results

When target is much harder
to extract than spurious...

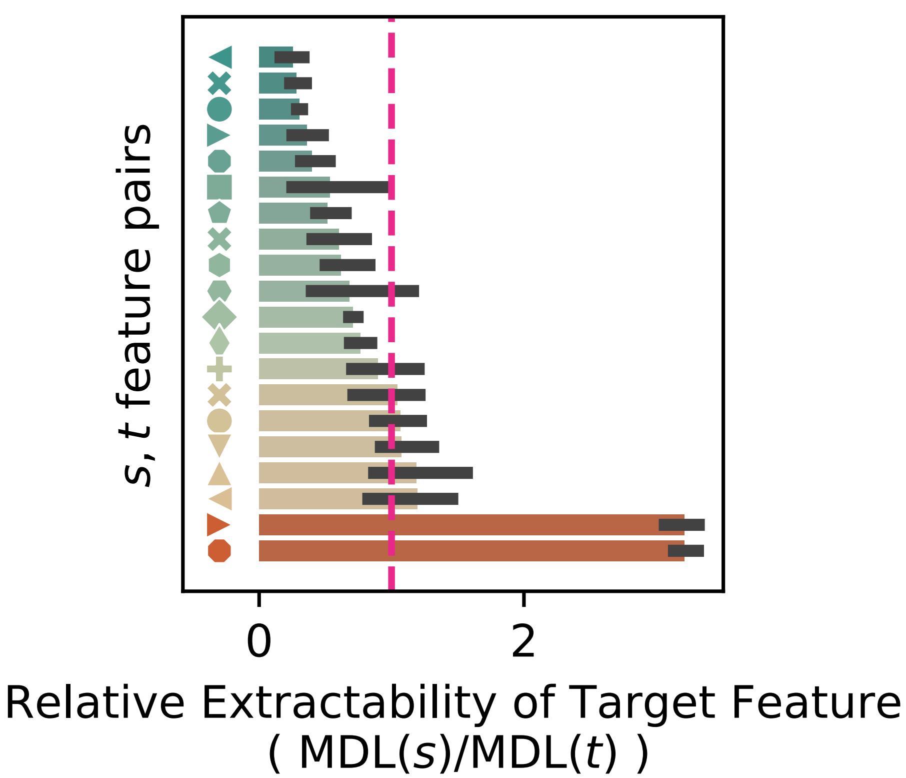


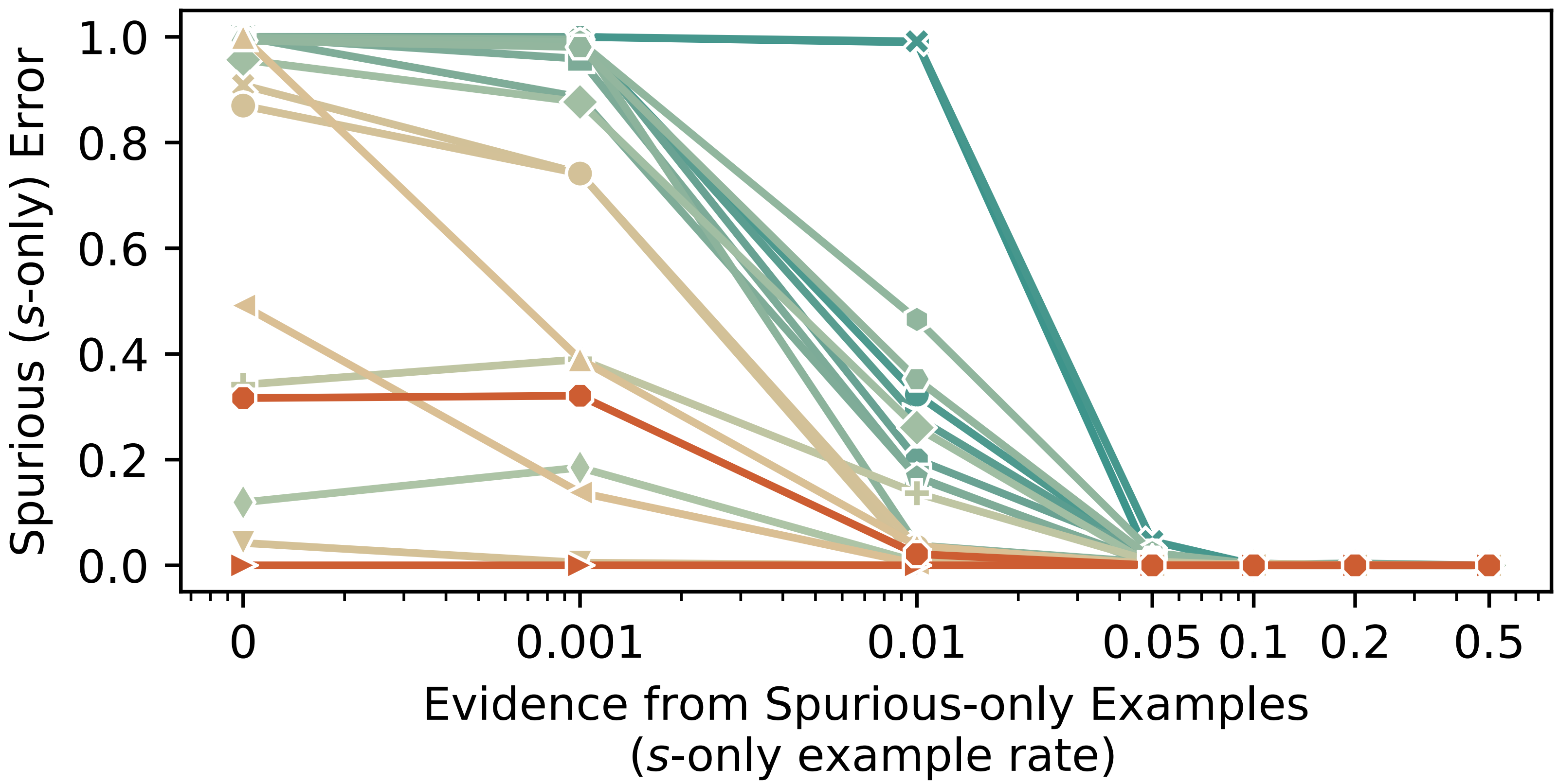


...model requires substantial training
incentive (e.g., 5% of training examples).

Jha, Lovering, Linzen, and Pavlick (2020)

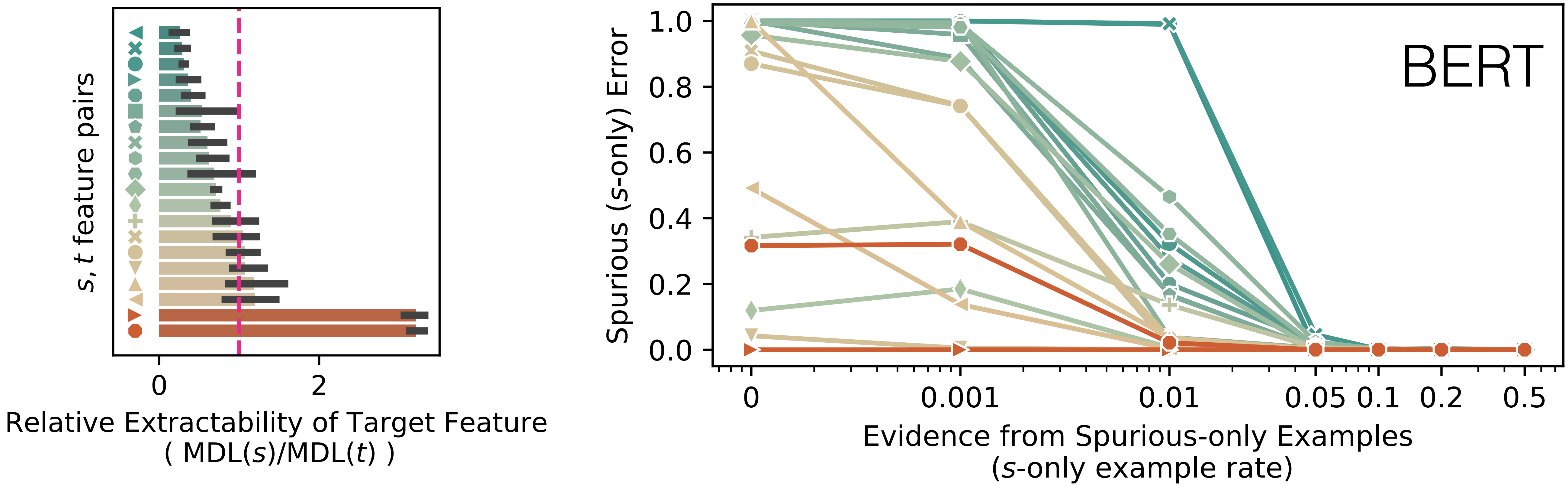
Results

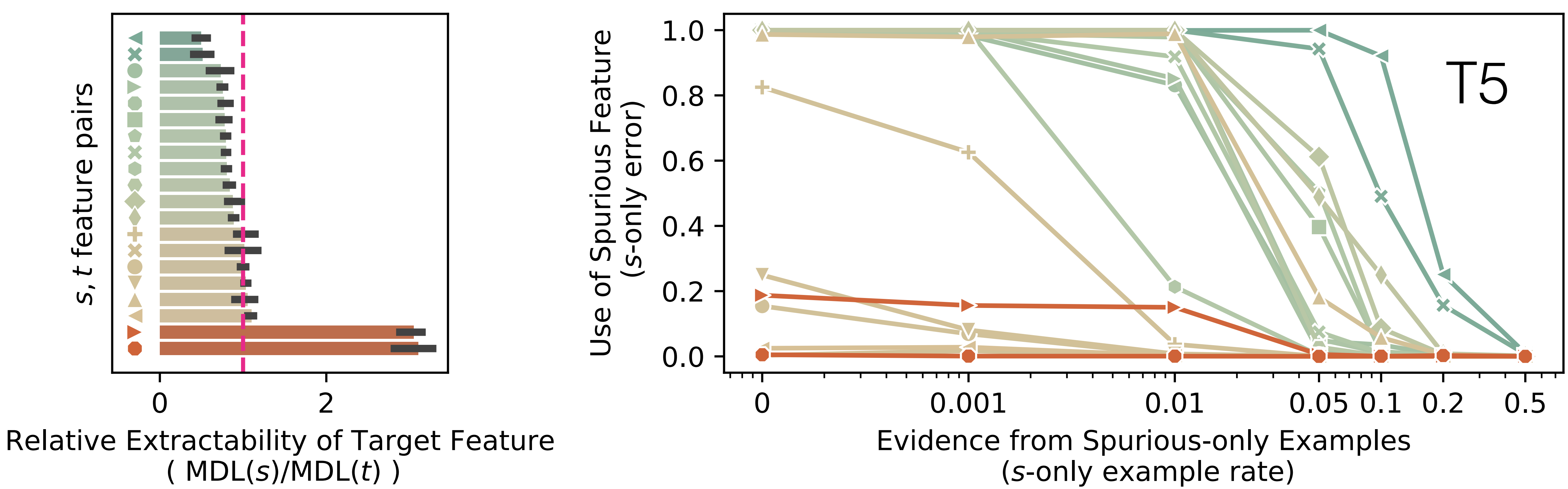




In general, learning curves track order predicted by relative MDL metric

Information-Theoretic Probing Explains Reliance on Spurious Heuristics
Jha, Lovering, Linzen, and Pavlick (2020)

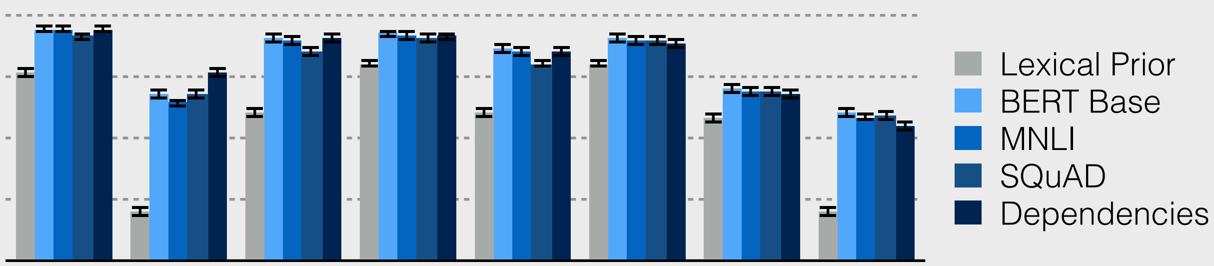




Jha, Lovering, Linzen, and Pavlick (2020)

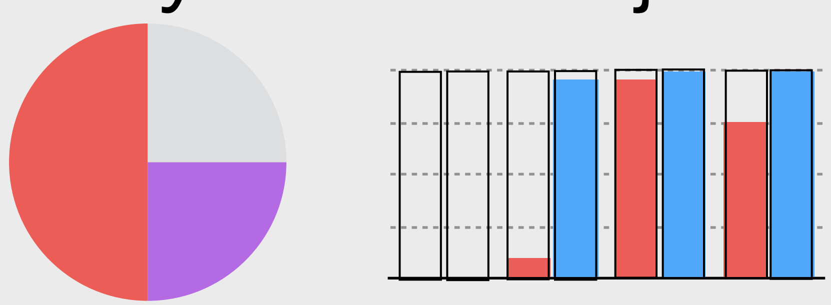
Linguistic features seem to be "there" after pretraining, but fine-tuned models don't use them… why?

~~Maybe the features are erased during finetuning?~~



No obvious drop in probing accuracy after fine-tuning.

~~Maybe there just isn't enough signal in training?~~

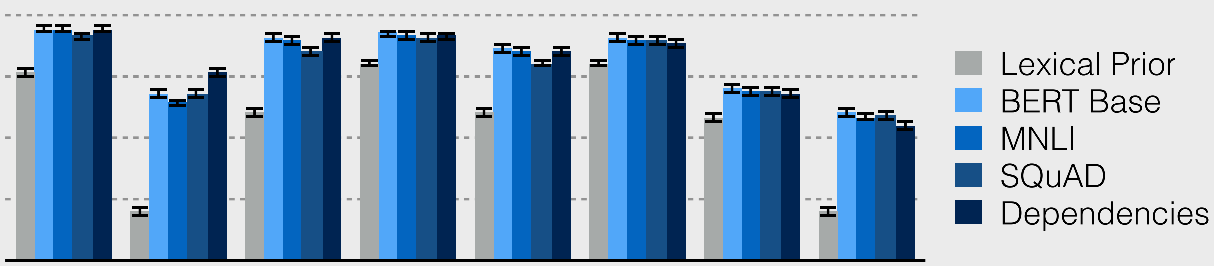


Different features behave differently given the same training data.

Maybe it's not just a matter of features being "there" or "not there"…?

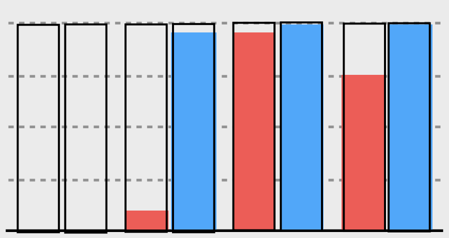
Linguistic features seem to be "there" after pretraining, but fine-tuned models don't use them… why?

~~Maybe the features are erased during finetuning?~~



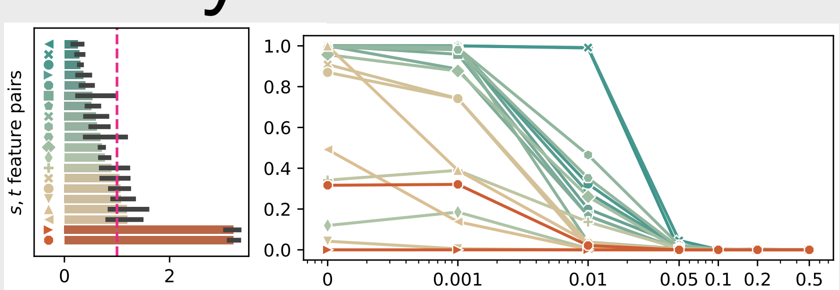
No obvious drop in probing accuracy after fine-tuning.

~~Maybe there just isn't enough signal in training?~~



Different features behave differently given the same training data.

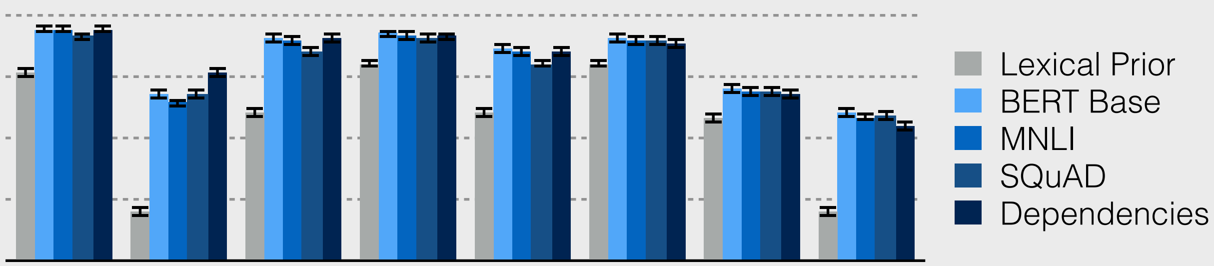
Maybe features aren't just "there" or "not there"?



Training data alone can't explain model behavior; models need little incentive when features are easy to extract.

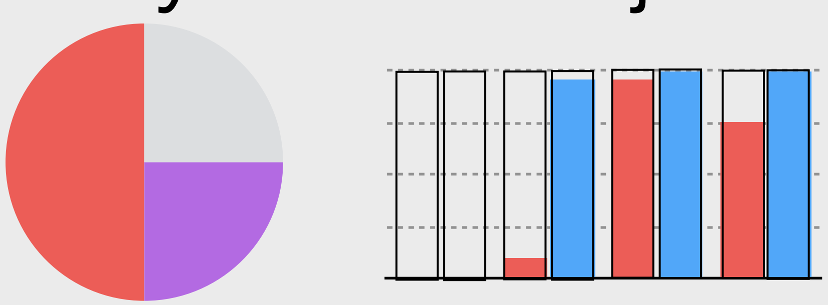
Linguistic features seem to be “there” after pretraining, but fine-tuned models don’t use them… why?

~~Maybe the features are erased during finetuning?~~



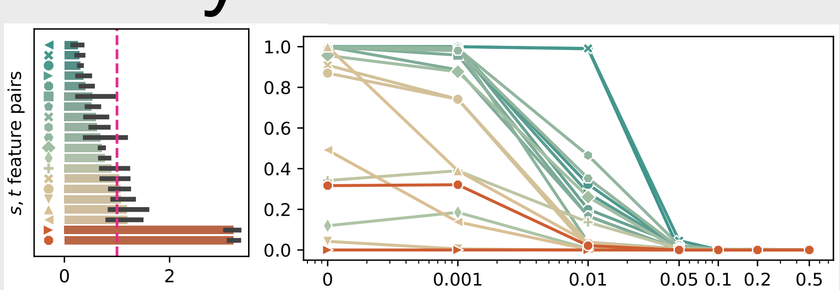
No obvious drop in probing accuracy after fine-tuning.

~~Maybe there just isn’t enough signal in training?~~



Different features behave differently given the same training data.

Maybe features aren’t just “there” or “not there”?



…so…?

Discussion

Discussion

- Takeaway: Pretraining can be viewed as endowing inductive biases, pushing the model to prefer certain solutions over others

Discussion

- Takeaway: Pretraining can be viewed as endowing inductive biases, pushing the model to prefer certain solutions over others
- Takeaway: To improve model behavior, we can tweak training data or tweak pretraining (ideally in principled ways). We might not always have control over both

Discussion

- Takeaway: Pretraining can be viewed as endowing inductive biases, pushing the model to prefer certain solutions over others

- Takeaway: To improve model behavior, we can tweak training data or tweak pretraining (ideally in principled ways). We might not always have control over both (Personally, I like to assume we can't chose our training data…)

Discussion

- Takeaway: Pretraining can be viewed as endowing inductive biases, pushing the model to prefer certain solutions over others
- Takeaway: To improve model behavior, we can tweak training data or tweak pretraining (ideally in principled ways). We might not always have control over both (Personally, I like to assume we can't chose our training data…)
- Implications: Innate structure via distributional pretraining? A happy solution to poverty of the stimulus that everyone can get behind? ;)

Discussion

- Takeaway: Pretraining can be viewed as endowing inductive biases, pushing the model to prefer certain solutions over others
- Takeaway: To improve model behavior, we can tweak training data or tweak pretraining (ideally in principled ways). We might not always have control over both (Personally, I like to assume we can't chose our training data…)
- Implications: Innate structure via distributional pretraining? A happy solution to poverty of the stimulus that everyone can get behind? ;)
- Implications: Innate structure from non-language pre-training? E.g., objects and agents by modeling the physical world?

Thank you!