New faithfulness-Centric Interpretability Paradigms for Natural Language Processing

Andreas Madsen

Outline

1. Background on Interpretability

New Interpretability Paradigms

2. Faithfulness Measure Models

3. Self-explanations

Interpretability

"The ability to explain or present
(a model or dataset)
in understandable terms
to a human."

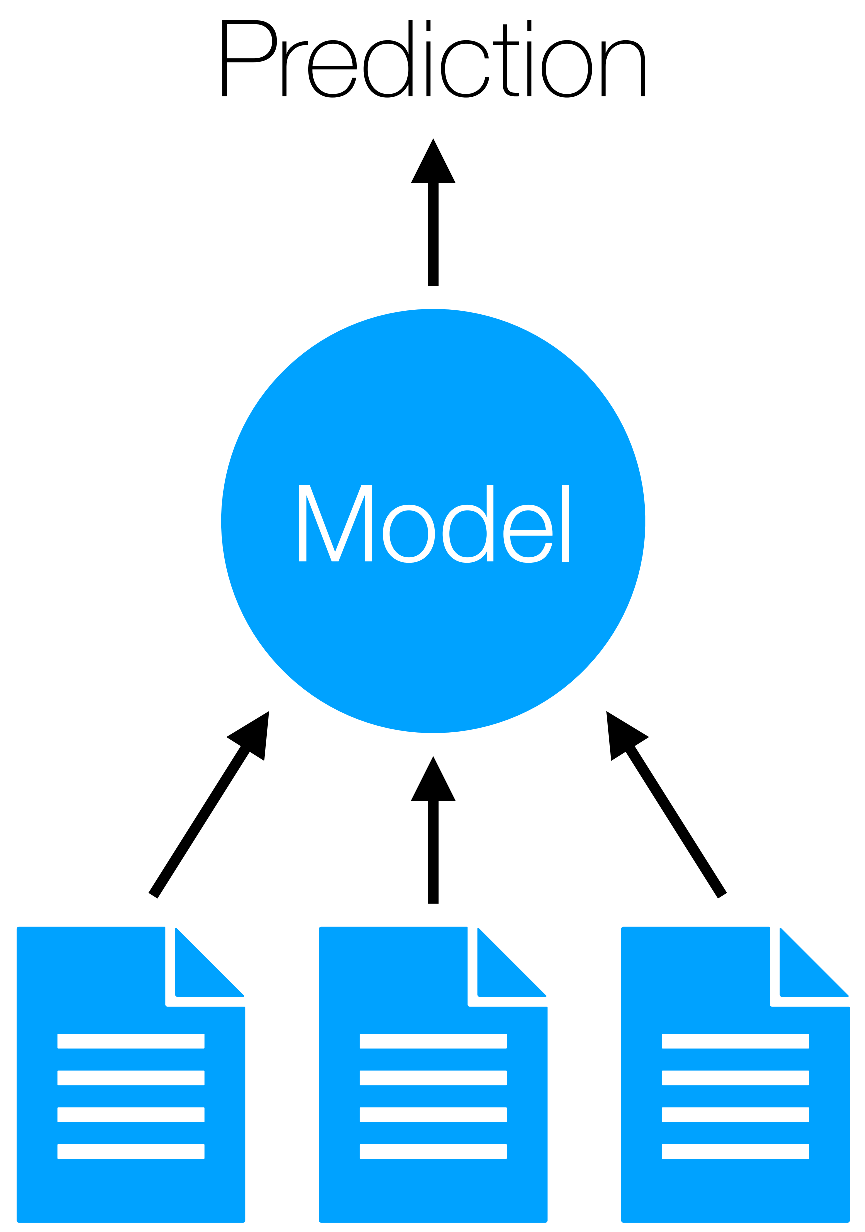
Doshi-Velez, F., & Kim, B (2017).
Towards A Rigorous Science of Interpretable Machine Learning.

Use cases

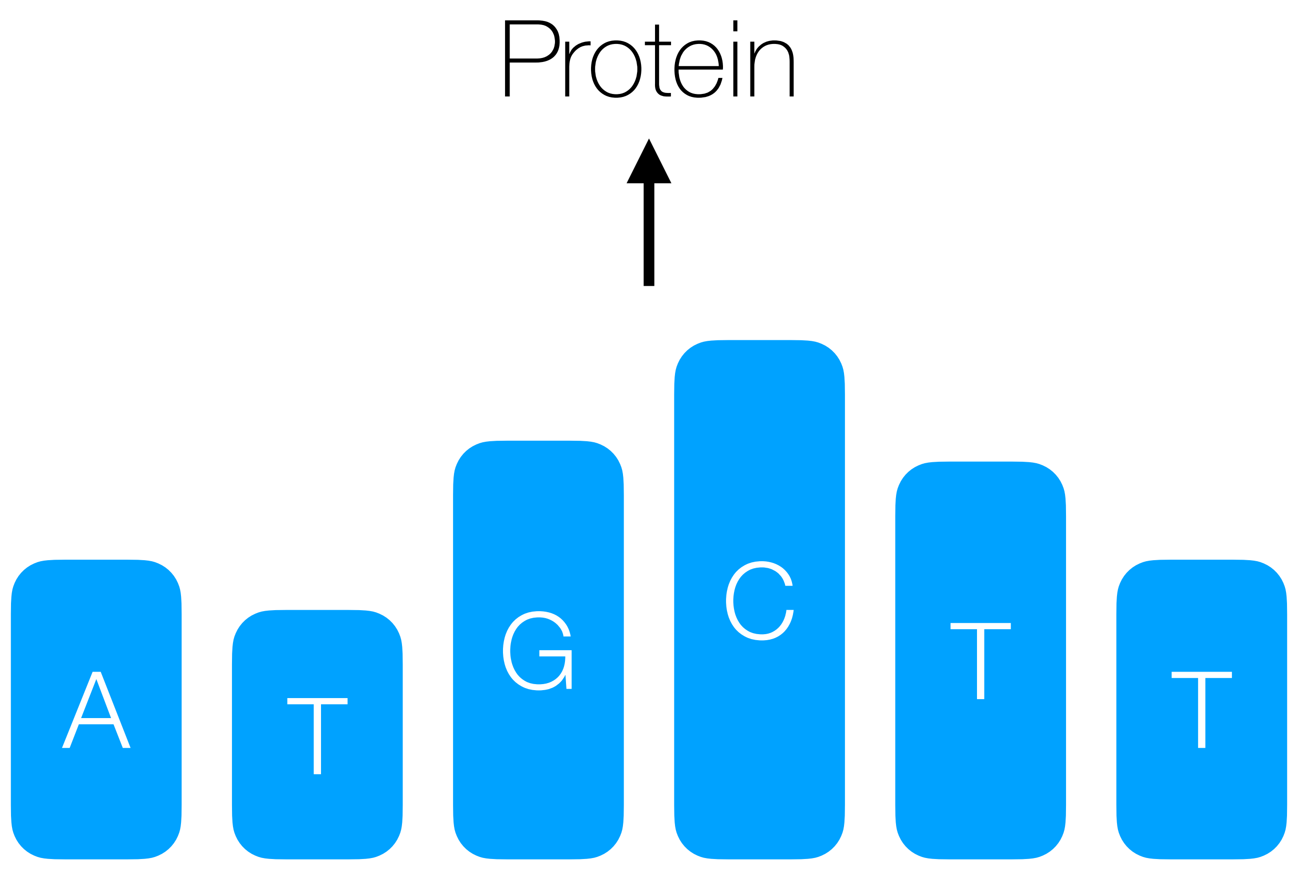
Identify model issues

Only people with a CS degree are qualified typists [1].

Identify actionable fixes

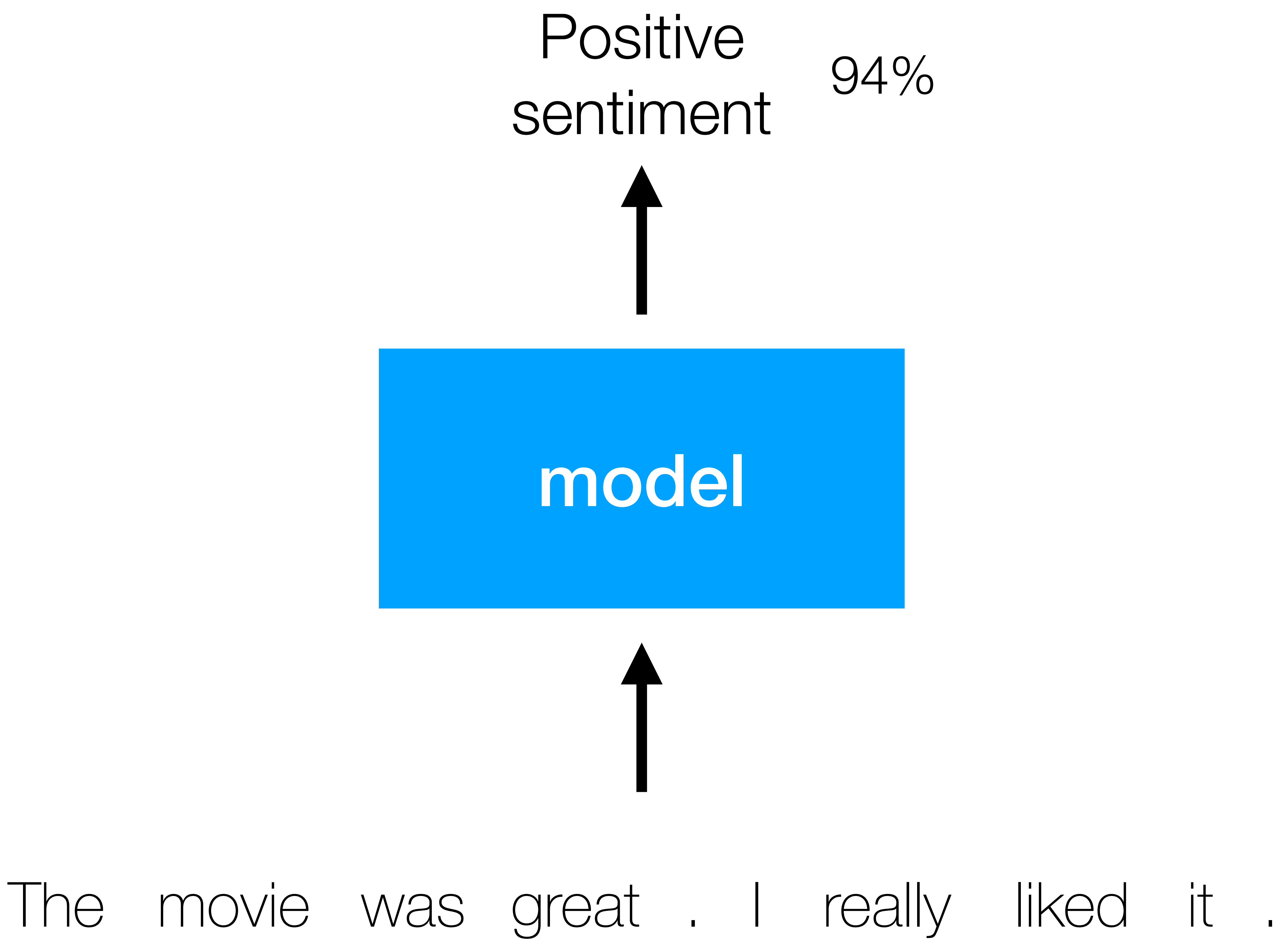


Scientific discovery

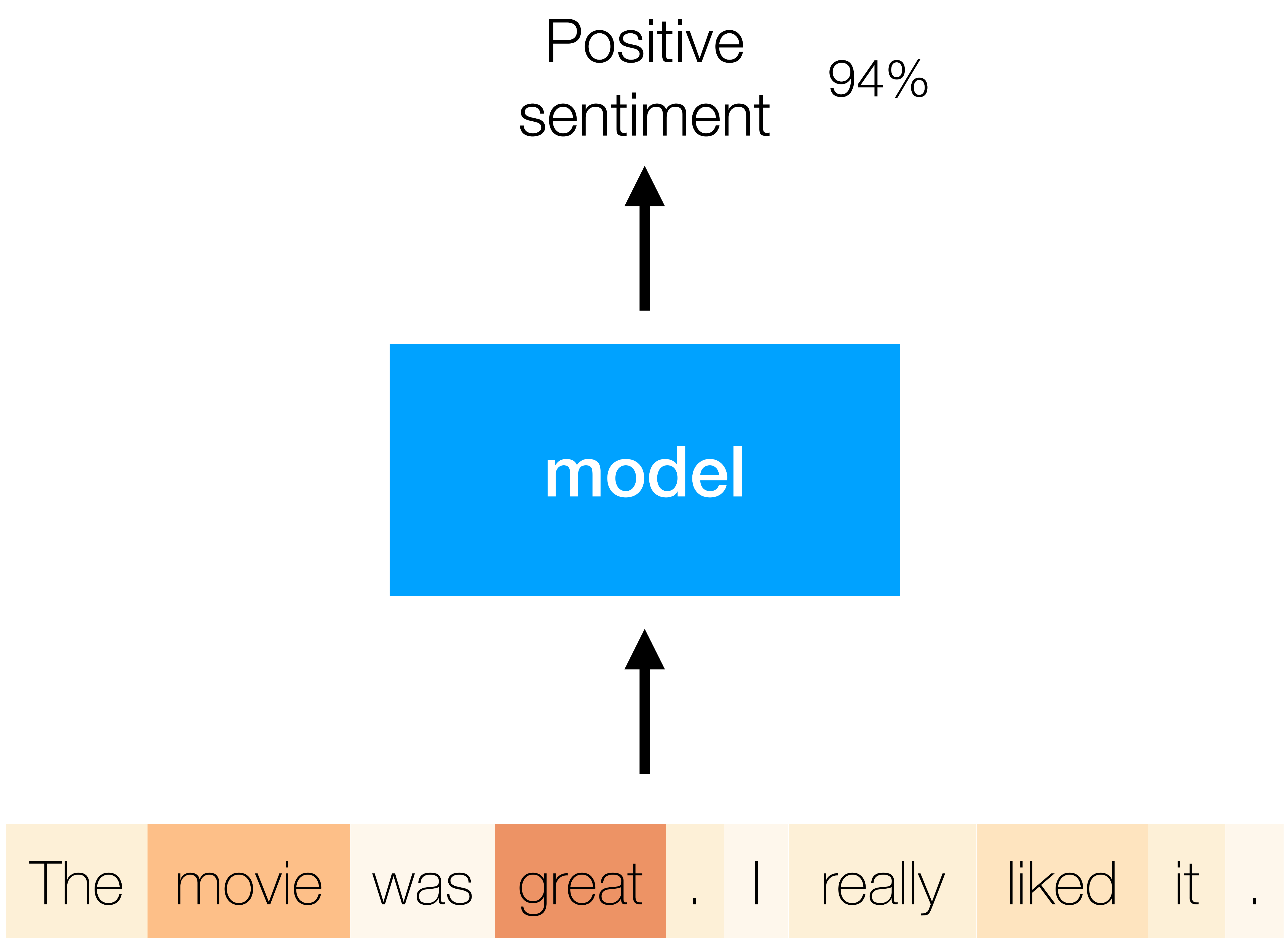


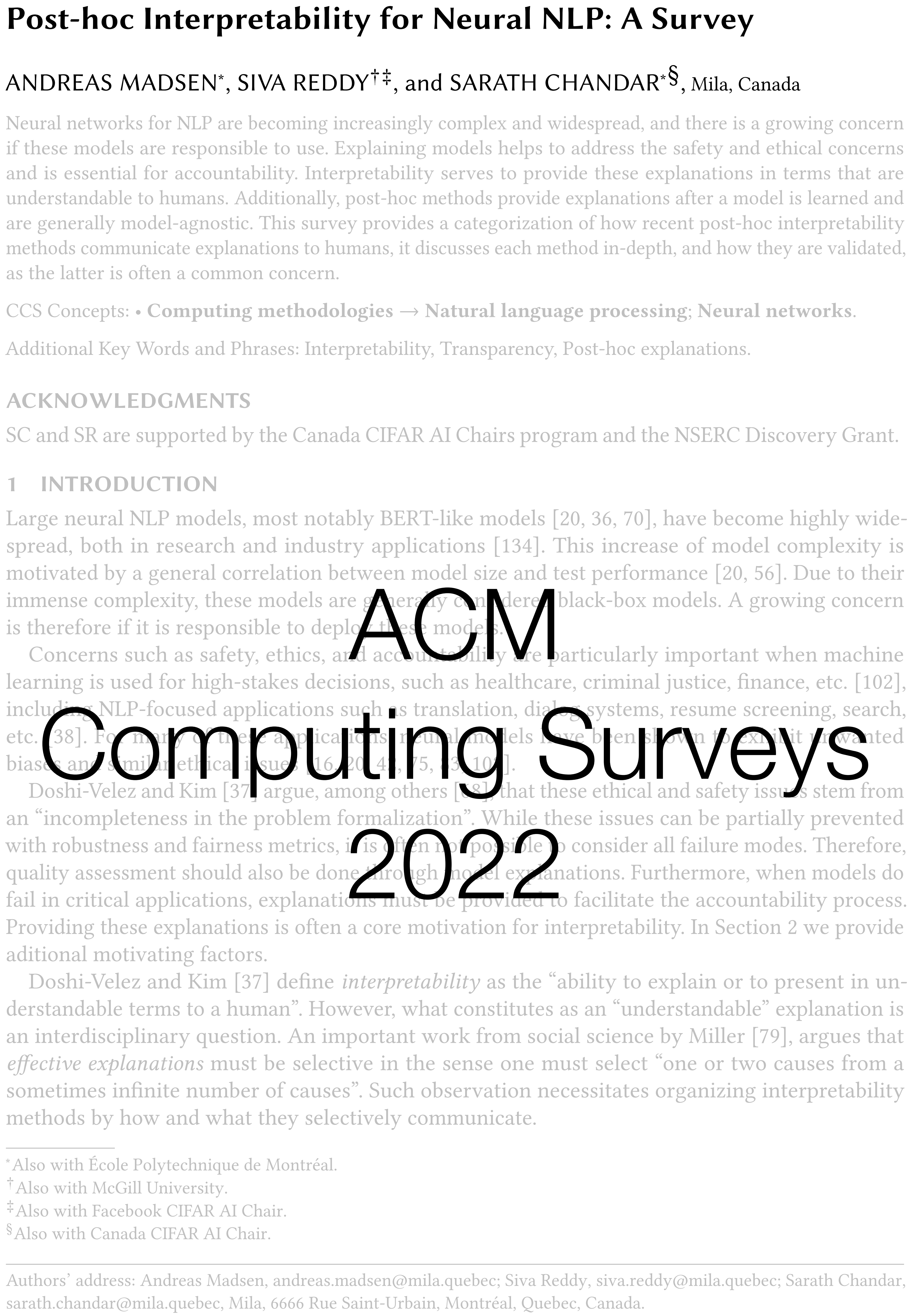
[1] Fuller, J. (2021). Companies Need More Workers. Why Do They Reject Millions of Résumés? The Project on Workforce.

Sentiment Classification



Importance Measures





Post-hoc Interpretability for Neural NLP: A Survey

ANDREAS MADSEN*, SIVA REDDY†‡, and SARATH CHANDAR*§, Mila, Canada

Neural networks for NLP are becoming increasingly complex and widespread, and there is a growing concern if these models are responsible to use. Explaining models helps to address the safety and ethical concerns and is essential for accountability. Interpretability serves to provide these explanations in terms that are understandable to humans. Additionally, post-hoc methods provide explanations after a model is learned and are generally model-agnostic. This survey provides a categorization of how recent post-hoc interpretability methods communicate explanations to humans, it discusses each method in-depth, and how they are validated, as the latter is often a common concern.

CCS Concepts: • **Computing methodologies → Natural language processing**; **Neural networks**.

Additional Key Words and Phrases: Interpretability, Transparency, Post-hoc explanations.

ACKNOWLEDGMENTS

SC and SR are supported by the Canada CIFAR AI Chairs program and the NSERC Discovery Grant.

1 INTRODUCTION

Large neural NLP models, most notably BERT-like models [20, 36, 70], have become highly widespread, both in research and industry applications [134]. This increase of model complexity is motivated by a general correlation between model size and test performance [20, 56]. Due to their immense complexity, these models are g[illegible] c[illegible]dered black-box models. A growing concern is therefore if it is responsible to depl[illegible] models.

Concerns such as safety, ethics, a[illegible] acc[illegible]ability are particularly important when machine learning is used for high-stakes decisions, such as healthcare, criminal justice, finance, etc. [102], includ[illegible] NLP-focused applications such as translation, dia[illegible] systems, resume screening, search, etc. [38]. [illegible] models have been shown to exhibit unwanted bias[illegible] [illegible].

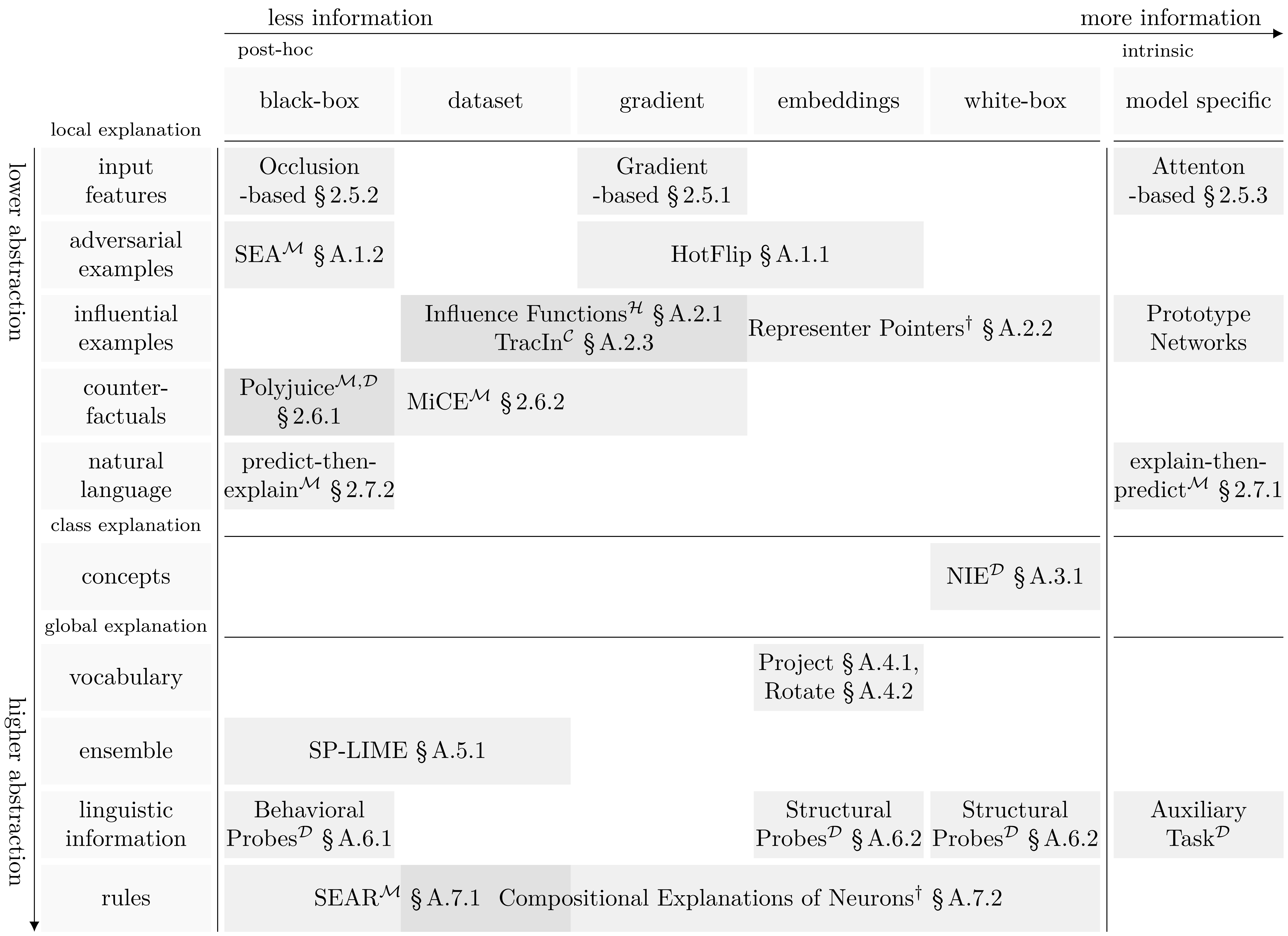
Doshi-Velez and Kim [37] argue, among others [illegible], that these ethical and safety issues stem from an “incompleteness in the problem formalization”. While these issues can be partially prevented with robustness and fairness metrics, it is often not possible to consider all failure modes. Therefore, quality assessment should also be done through model explanations. Furthermore, when models do fail in critical applications, explanations must be provided to facilitate the accountability process. Providing these explanations is often a core motivation for interpretability. In Section 2 we provide aditional motivating factors.

Doshi-Velez and Kim [37] define *interpretability* as the “ability to explain or to present in understandable terms to a human”. However, what constitutes as an “understandable” explanation is an interdisciplinary question. An important work from social science by Miller [79], argues that *effective explanations* must be selective in the sense one must select “one or two causes from a sometimes infinite number of causes”. Such observation necessitates organizing interpretability methods by how and what they selectively communicate.

*Also with École Polytechnique de Montréal.
†Also with McGill University.
‡Also with Facebook CIFAR AI Chair.
§Also with Canada CIFAR AI Chair.

Authors' address: Andreas Madsen, andreas.madsen@mila.quebec; Siva Reddy, siva.reddy@mila.quebec; Sarath Chandar, sarath.chandar@mila.quebec, Mila, 6666 Rue Saint-Urbain, Montréal, Quebec, Canada.

ACM
Computing Surveys
2022



Post-hoc Interpretability for Neural NLP: A Survey

ANDREAS MADSEN*, SIVA REDDY†‡, and SARATH CHANDAR*§, Mila, Canada

Neural networks for NLP are becoming increasingly complex and widespread, and there is a growing concern if these models are responsible to use. Explaining models helps to address the safety and ethical concerns and is essential for accountability. Interpretability serves to provide these explanations in terms that are understandable to humans. Additionally, post-hoc methods provide explanations after a model is learned and are generally model-agnostic. This survey provides a categorization of how recent post-hoc interpretability methods communicate explanations to humans, it discusses each method in-depth, and how they are validated, as the latter is often a common concern.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; **Neural networks**.

Additional Key Words and Phrases: Interpretability, Transparency, Post-hoc explanations.

ACKNOWLEDGMENTS

SC and SR are supported by the Canada CIFAR AI Chairs program and the NSERC Discovery Grant.

1 INTRODUCTION

Large neural NLP models, most notably BERT-like models [20, 36, 70], have become highly widespread, both in research and industry applications [134]. This increase of model complexity is motivated by a general correlation between model size and test performance [20, 56]. Due to their immense complexity, these models are generally considered black-box models. A growing concern is therefore if it is responsible to deploy these models.

Concerns such as safety, ethics, and accountability are particularly important when machine learning is used for high-stakes decisions, such as healthcare, criminal justice, finance, etc. [102], including NLP-focused applications such as translation, dialog systems, resume screening, search, etc. [38]. For many of these applications, neural models have been shown to exhibit unwanted biases and similar ethical issues [6, 20, 42, 75, 89, 106].

Doshi-Velez and Kim [37] argue, among others [78], that these ethical and safety issues stem from an "incompleteness in the problem formalization". While these issues can be partially prevented with robustness and fairness metrics, it is often not possible to consider all failure modes. Therefore, quality assessment should also be done through model explanations. Furthermore, when models do fail in critical applications, explanations must be provided to facilitate the accountability process. Providing these explanations is often a core motivation for interpretability. In Section 2 we provide aditional motivating factors.

Doshi-Velez and Kim [37] define *interpretability* as the "ability to explain or to present in understandable terms to a human". However, what constitutes as an "understandable" explanation is an interdisciplinary question. An important work from social science by Miller [79], argues that *effective explanations* must be selective in the sense one must select "one or two causes from a sometimes infinite number of causes". Such observation necessitates organizing interpretability methods by how and what they selectively communicate.

*Also with École Polytechnique de Montréal.
†Also with McGill University.
‡Also with Facebook CIFAR AI Chair.
§Also with Canada CIFAR AI Chair.

Authors' address: Andreas Madsen, andreas.madsen@mila.quebec; Siva Reddy, siva.reddy@mila.quebec; Sarath Chandar, sarath.chandar@mila.quebec, Mila, 6666 Rue Saint-Urbain, Montréal, Quebec, Canada.

ACM Computing Surveys 2022

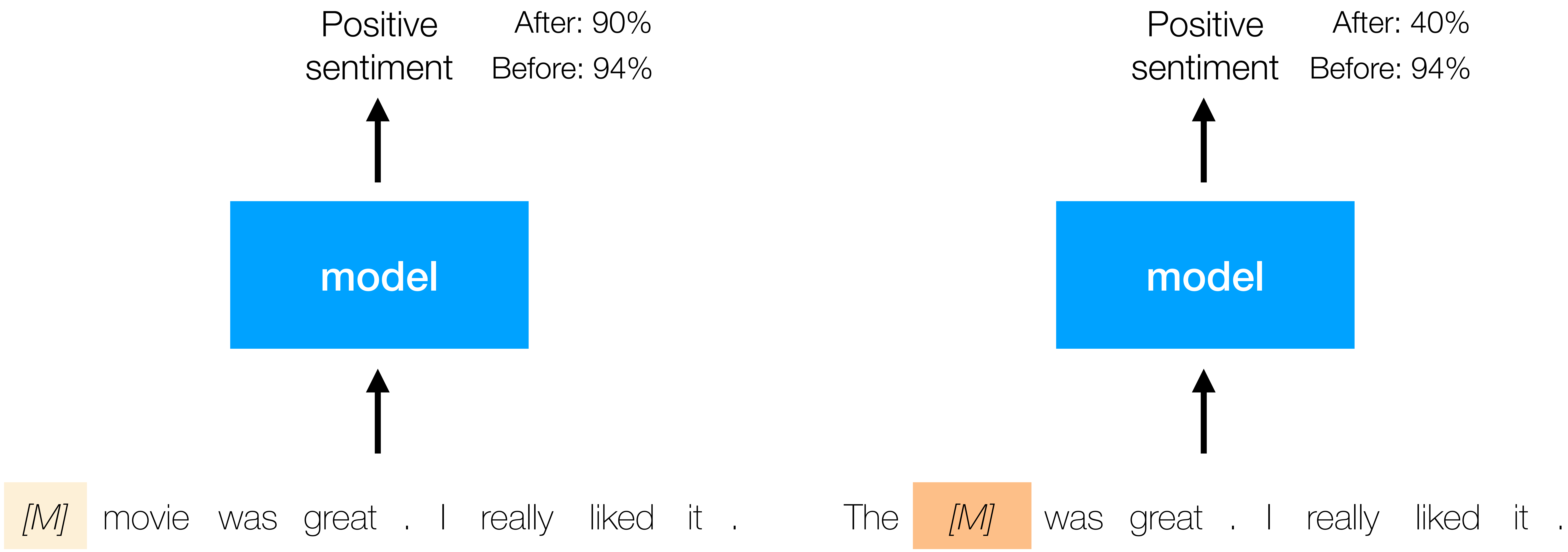
less information → more information

post-hoc | intrinsic

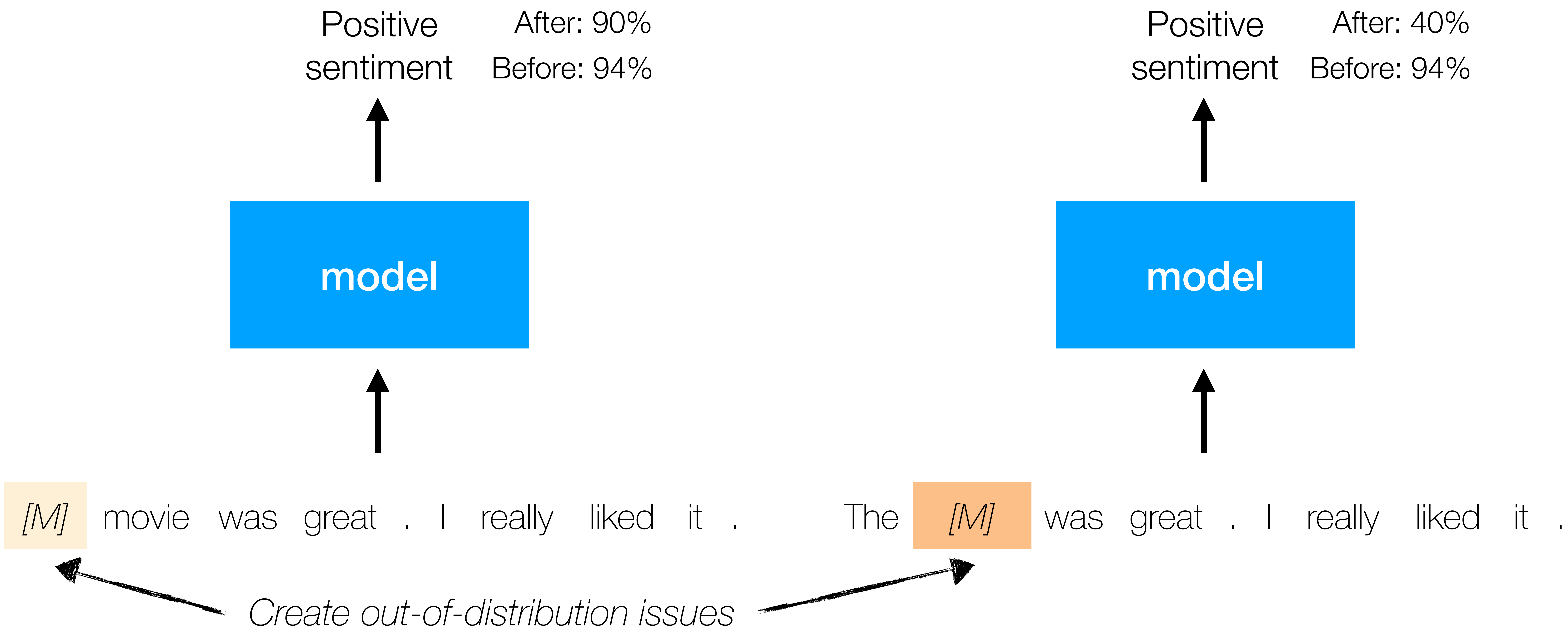
lower abstraction ↓ higher abstraction

	black-box	dataset	gradient	embeddings	white-box	model specific
local explanation						
input features	Occlusion -based § 2.5.2		Gradient -based § 2.5.1			Attenton -based § 2.5.3
adversarial examples	SEA$^{\mathcal{M}}$ § A.1.2		HotFlip § A.1.1			
influential examples		Influence Functions$^{\mathcal{H}}$ § A.2.1, TracIn$^{\mathcal{C}}$ § A.2.3	Representer Pointers† § A.2.2			Prototype Networks
counter-factuals	Polyjuice$^{\mathcal{M},\mathcal{D}}$ § 2.6.1	MiCE$^{\mathcal{M}}$ § 2.6.2				
natural language	predict-then-explain$^{\mathcal{M}}$ § 2.7.2					explain-then-predict$^{\mathcal{M}}$ § 2.7.1
class explanation						
concepts					NIE$^{\mathcal{D}}$ § A.3.1	
global explanation						
vocabulary				Project § A.4.1, Rotate § A.4.2		
ensemble	SP-LIME § A.5.1					
linguistic information	Behavioral Probes$^{\mathcal{D}}$ § A.6.1			Structural Probes$^{\mathcal{D}}$ § A.6.2	Structural Probes$^{\mathcal{D}}$ § A.6.2	Auxiliary Task$^{\mathcal{D}}$
rules	SEAR$^{\mathcal{M}}$ § A.7.1	Compositional Explanations of Neurons† § A.7.2				

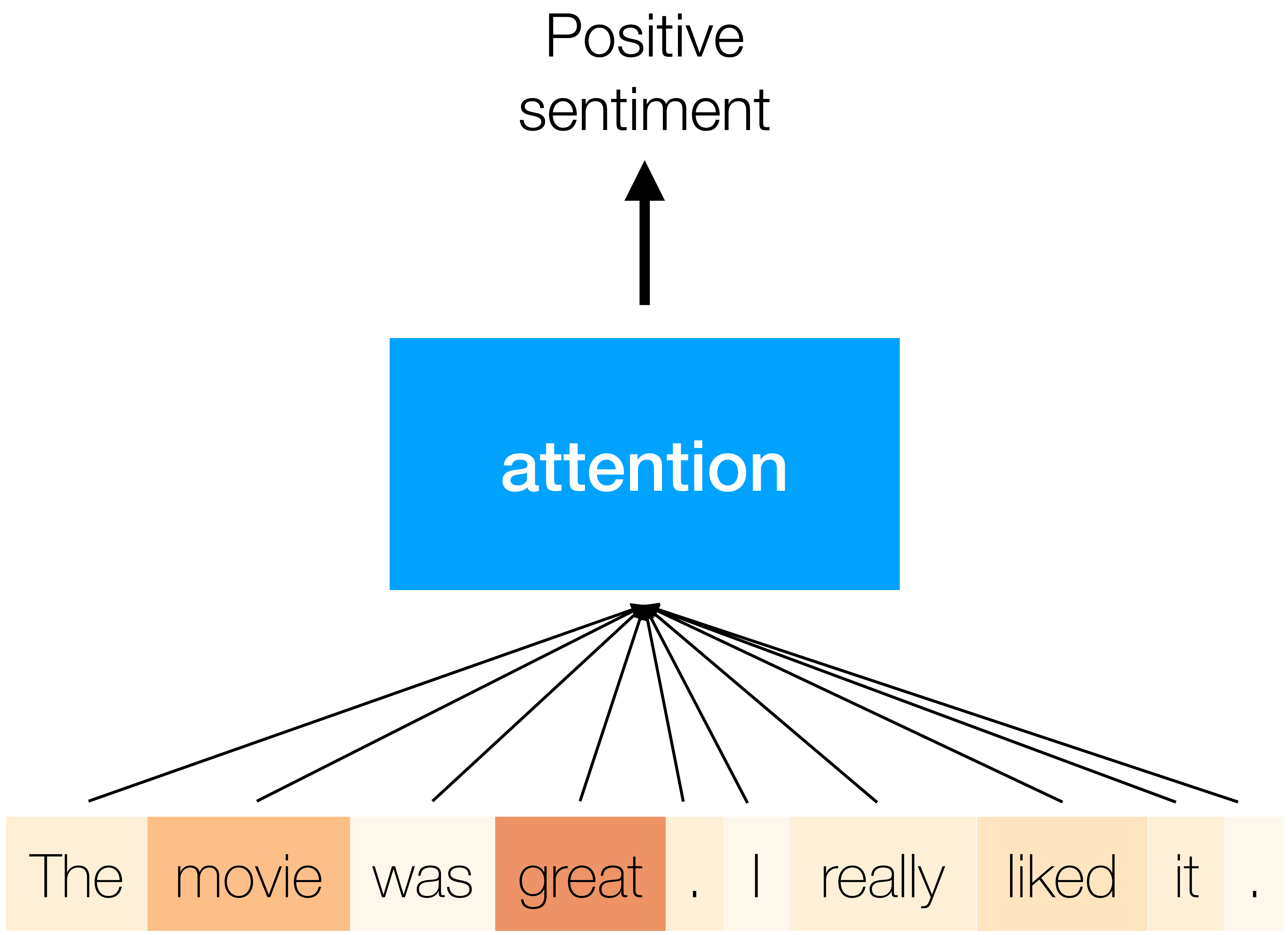
Leave one out (LOO)



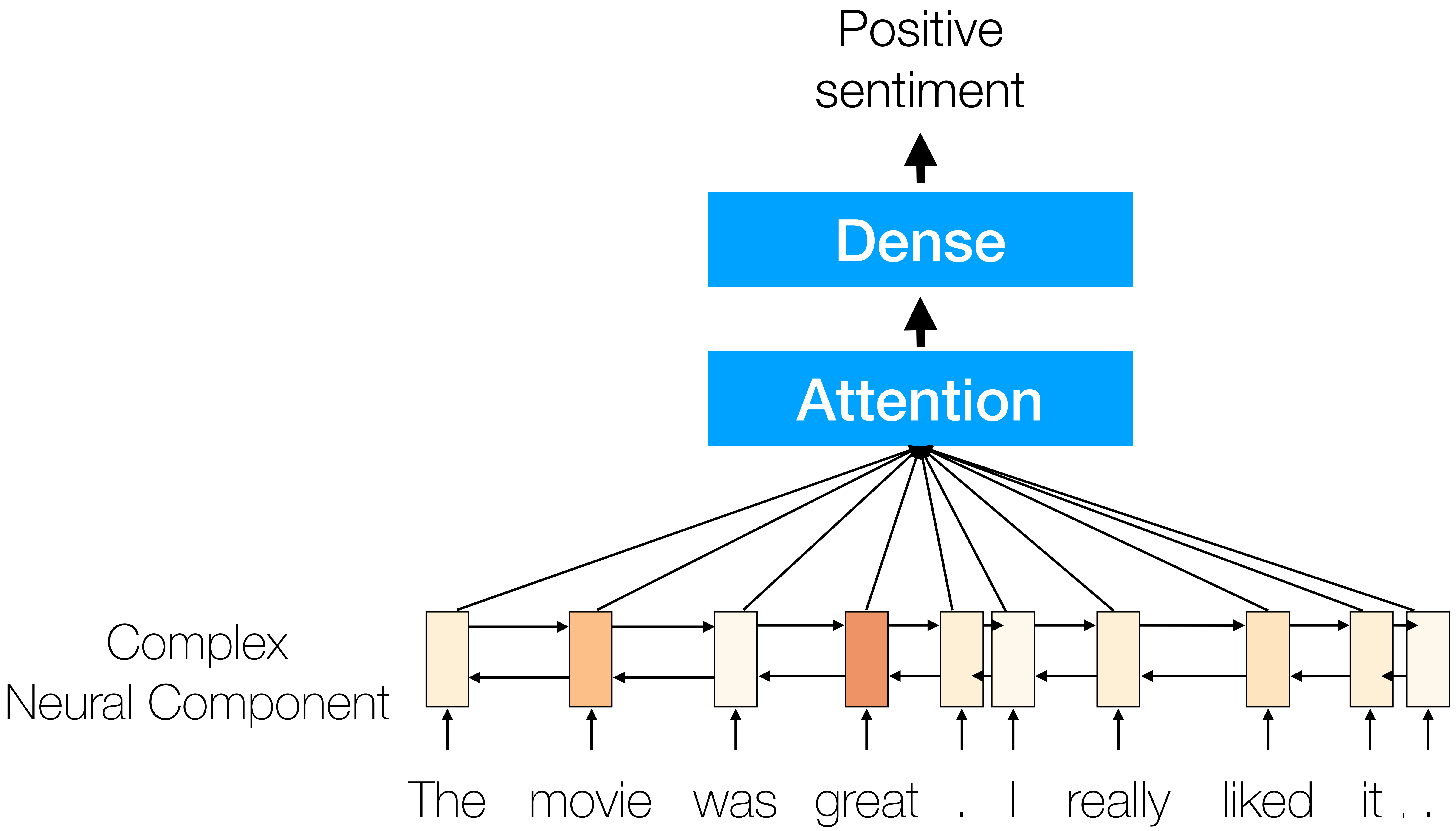
Leave one out (LOO)



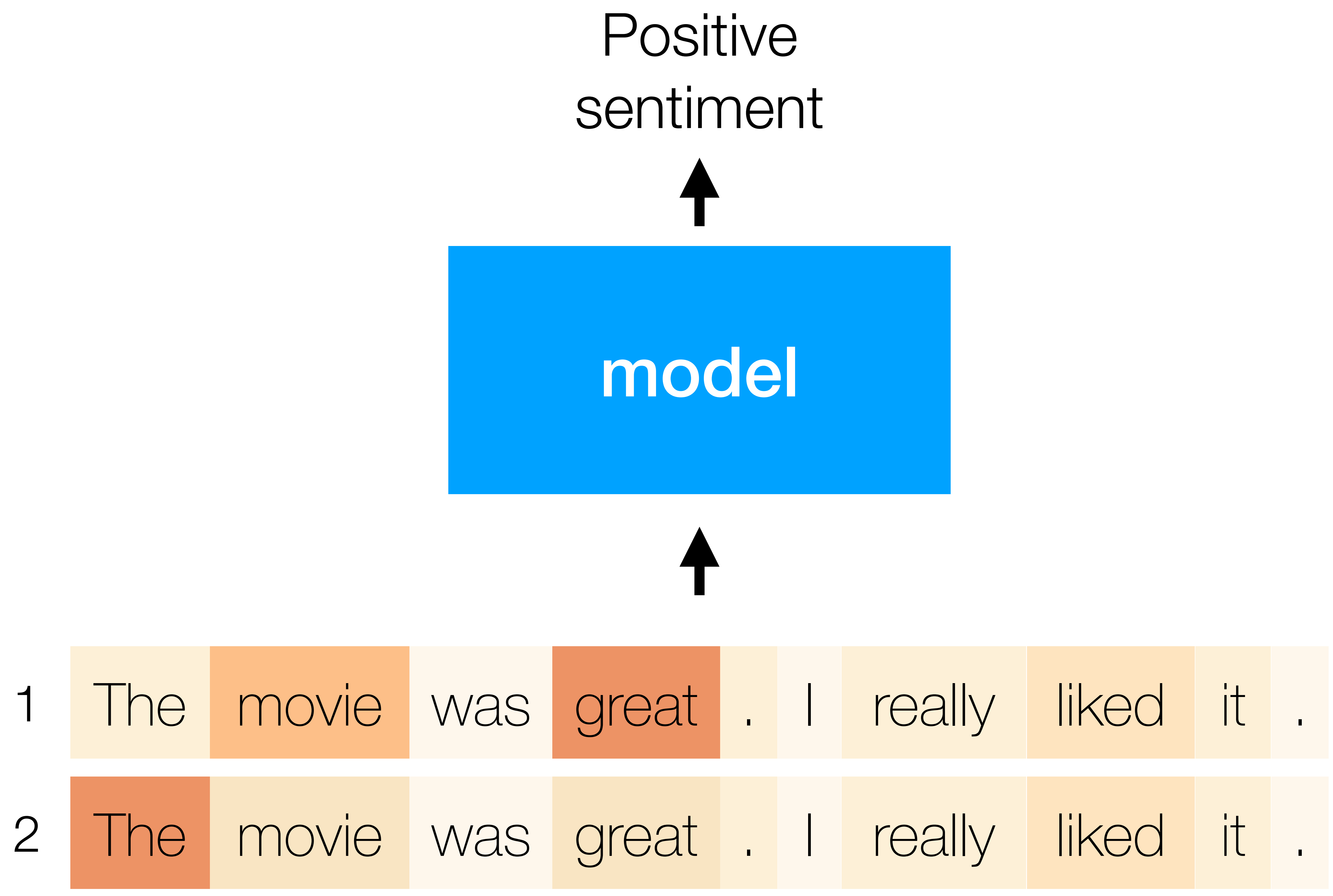
Attention


Positive
sentiment
attention
The
movie
was
great
.
I
really
liked
it
.

Attention


Positive
sentiment
Dense
Attention
Complex
Neural Component
The movie was great . I really liked it . .

Disagreement problem



Desirables

faithfulness [1]

"How accurately it (the explanation) reflects the true reasoning process of the model."

human-groundedness [2]

How useful is the explanation to humans.

[1] Jacovi, A., & Goldberg, Y. (2020). Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? ACL 2020

[2] Doshi-Velez, F., & Kim, B (2017). Towards A Rigorous Science of Interpretable Machine Learning.

Human-groundedness

Counterfactual generation

	$\mathbf{x}$	$p(y\|\mathbf{x};\theta)$	y
$\mathbf{x}$	the year 's best and most unpredictable comedy	0.91	pos
	the year 's worst and most unpredictable comedy	0.59	-
$\tilde{\mathbf{x}}$	the year 's worst and most predictable comedy	0.04	-
$\mathbf{x}$	we never feel anything for these characters	0.95	neg
	we can feel anything for these characters	0.73	-
$\tilde{\mathbf{x}}$	we can feel anything for these animals	0.01	-

Ross, A., Marasović, A., & Peters, M. (2021). Explaining NLP Models via Minimal Contrastive Editing (MiCE). Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021

Contrastive explanations

Input: *Can you stop the dog from*
Output: barking

1. Why did the model predict "barking"?
Can you stop the dog from

2. Why did the model predict "barking" ***instead of*** **"crying"?**
Can you stop the dog from

3. Why did the model predict "barking" ***instead of*** **"walking"?**
Can you stop the dog from

Yin, K., & Neubig, G. (2022). Interpreting Language Models with Contrastive Explanations. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing.

When are explanations faithful?

post-hoc

intrinsic

When are explanations faithful?

post-hoc

intrinsic

Interpretability is considered after the model is trained.

Leave-one-out, Gradient-based

When are explanations faithful?

post-hoc

Interpretability is considered after the model is trained.

Leave-one-out, Gradient-based

intrinsic

Models are architecturally constrained to be explained.

Attention, Decision Trees

When are explanations faithful?

post-hoc	intrinsic
Interpretability is considered after the model is trained.	Models are architecturally constrained to be explained.
	Only models designed to be explained can be explained.

When are explanations faithful?

post-hoc

intrinsic

Models are architecturally constrained to be explained.

Only models designed to be explained can be explained.

When are explanations faithful?

post-hoc

intrinsic

Models are architecturally constrained to be explained.

Only models designed to be explained can be explained.

Intrinsic models can have high-performance too.

When are explanations faithful?

post-hoc

intrinsic

Interpretability is considered after the model is trained.

Black-box models are more general purpose.

When are explanations faithful?

post-hoc

intrinsic

Interpretability is considered after the model is trained.

Any model can be explained.

Black-box models are more general purpose.

The evolution of paradigms

Light is a particle.

Light is a wave.

The evolution of paradigms

Quantum mechanisms

Light is a particle.	Light is a wave.

post-hoc

intrinsic

Only models designed to be explained can be explained.

Black-box models are more general purpose.

New Interpretability Paradigms

Only models designed to be explained can be explained.

Black-box models are more general purpose.

AI Interpretability Needs a New Paradigm

Andreas Madsen*
andreas.madsen@mila.quebec
Mila
Montréal, Quebec, Canada

Himabindu Lakkaraju
hlakkaraju@hbs.edu
Harvard University
Cambridge, Massachusetts, United States

Siva Reddy†‡
siva.reddy@mila.quebec
Mila
Montréal, Quebec, Canada

Sarath Chandar*§
sarath.chandar@mila.quebec
Mila
Montréal, Quebec, Canada

Abstract

Interpretability is the study of explaining models in understandable terms to humans. At present, interpretability is divided into two paradigms: the intrinsic paradigm, which believes that only models designed to be explained can be explained, and the post-hoc paradigm, which believes that black-box models can be explained. At the core of this debate is how each paradigm ensures its explanations are *faithful*, i.e., true to the model's behavior. This is important, as false but convincing explanations lead to unsupported confidence in artificial intelligence (AI), which can be dangerous. This article's perspective is that we should think about new paradigms while staying vigilant regarding faithfulness. First, by examining the history of paradigms in science, we see that paradigms are constantly evolving. Then, by examining the current paradigms, we can understand their underlying beliefs, the value they bring, and their limitations. Finally, this article presents 3 emerging paradigms for interpretability. The first paradigm designs models such that faithfulness can be easily measured. Another optimizes models such that explanations become faithful. The last paradigm proposes to develop models that produce both a prediction and an explanation.

Keywords

Interpretability, Explanations, Transparency, Paradigms, Post-hoc, Intrinsic, Ethics, Future work, Faithfulness measurable models, Self-explanations, Self-explaining models

ACM Reference Format:
Andreas Madsen, Himabindu Lakkaraju, Siva Reddy, and Sarath Chandar. 2024. AI Interpretability Needs a New Paradigm. In *Proceedings of Communications of the ACM (CACM)*. ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

There was a time in physics, in the late 17th century, when Isaac Newton insisted that light is a particle and Christiaan Huygens insisted that light is a wave. These ideas were seemingly irreconcilable at the time. Of course, now we have a much better theory, and we understand that light can be seen as both a wave and a particle.[1]

In 1874, Georg Cantor proposed set theory and showed there exists at least two kinds of infinity. This divided the mathematical field. The Intuitionists, who named Cantor's theory nonsense, thought that math was a pure creation of the mind and that these infinities

How to provide and ensure faithful explanations for complex general-purpose neural NLP models?

Research question

This question can be answered:

- By developing **new paradigms** that design models to be explained without employing architectural constraints.
- By focusing on developing **accurate faithfulness metrics**.
- By focusing on **importance measures** that have had a notoriously troubling history regarding faithfulness.
- By taking advantage of properties specific to natural language and **NLP** models.

Research hypothesis

Potential paradigms

Faithfulness measurable models

Model is designed such that measuring faithfulness is easy.

ICML 2024
Spotlight

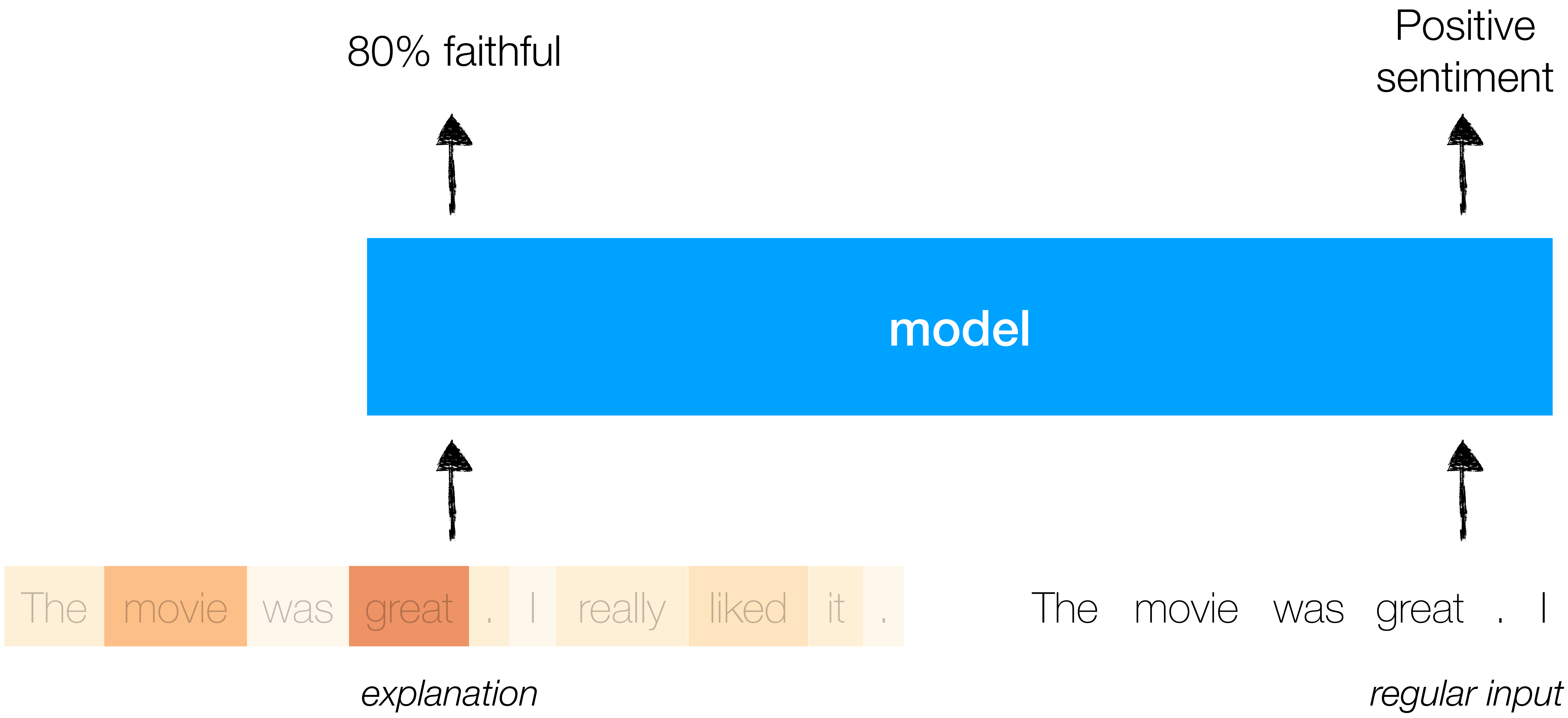
Self-explanations

Model is designed such that it can explain itself.

ACL 2024
Findings

Faithfulness
measurable models

Faithfulness measurable model



erasure-metric

If a token is truly important,
then if the token is removed,
the model's prediction should
change significantly.

Samek, W., et al. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE 2017.*

Hooker, S., Erhan, D., Kindermans, P.-J. J., & Kim, B. A benchmark for interpretability methods in deep neural networks. NeurIPS 2019.

Evaluating the Faithfulness of Importance Measures in NLP by Recursively Masking Allegedly Important Tokens and Retraining

Andreas Madsen[1,2] **Nicholas Meade**[1,3,*] **Vaibhav Adlakha**[1,3,*] **Siva Reddy**[1,3,4]
[1] Mila – Quebec AI Institute [2] Polytechnique Montréal
[3] McGill University [4] Facebook CIFAR AI Chair
{firstname.lastname}@mila.quebec

EMNLP 2022

Findings

Abstract

To explain NLP models a popular approach is to use importance measures, such as attention, which inform input tokens are important for making a prediction. However, an open question is how well these explanations accurately reflect a model's logic, a property called *faithfulness*.

To answer this question, we propose Recursive ROAR, a new faithfulness metric. This works by recursively masking allegedly important tokens and then retraining the model. The principle is that this should result in worse model performance compared to masking random tokens. The result is a performance curve given a masking-ratio. Furthermore, we propose a summarizing metric using relative area-between-curves (RACU), which allows for easy comparison across papers, models, and tasks.

We evaluate 4 different importance measures on 8 different datasets, using both LSTM-attention models and RoBERTa models. We find that the faithfulness of importance measures is both model-dependent and task-

are relevant for a given prediction. This type of explanation is called an importance measure.

A major challenge in the field of interpretability is ensuring that an explanation is *faithful*: "a faithful interpretation is one that accurately represents the reasoning process behind the model's prediction" (Jacovi and Goldberg, 2020). Unfortunately, importance measures that are claimed to have strong theoretical foundations and are widely used in practice (Bhatt et al., 2019) often later turn out to be questionable (Hooker et al., 2019; Kindermans et al., 2019; Adebayo et al., 2018; Jain and Wallace, 2019; Wiegreffe and Pinter, 2019).

Accurately measuring if an explanation is faithful is therefore paramount. Such *faithfulness* metrics are difficult to develop as the models are too complex to know what the correct explanation is. Doshi-Velez and Kim (2017) says a *faithfulness* metric should use "some formal definition of interpretability as a proxy for explanation quality."

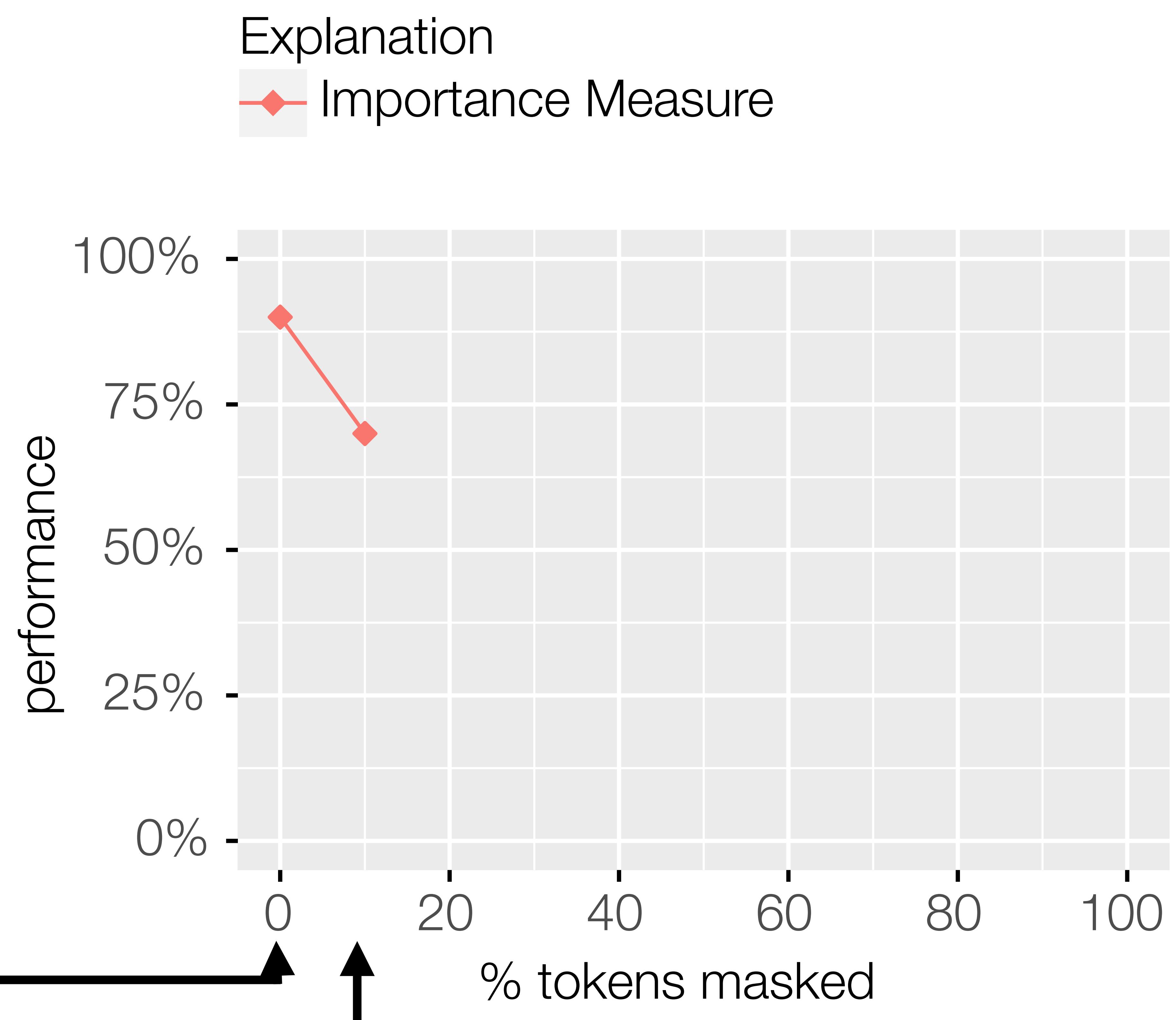
In this work, we use the definition of *faithfulness* by Samek et al. (2017) and Hooker et al. (2019): if information (input tokens) is truly important, then

ROAR

Compute importance measure

Repeat this:

1. Mask 10% more of the dataset
2. Retrain the model
3. Measure the performance



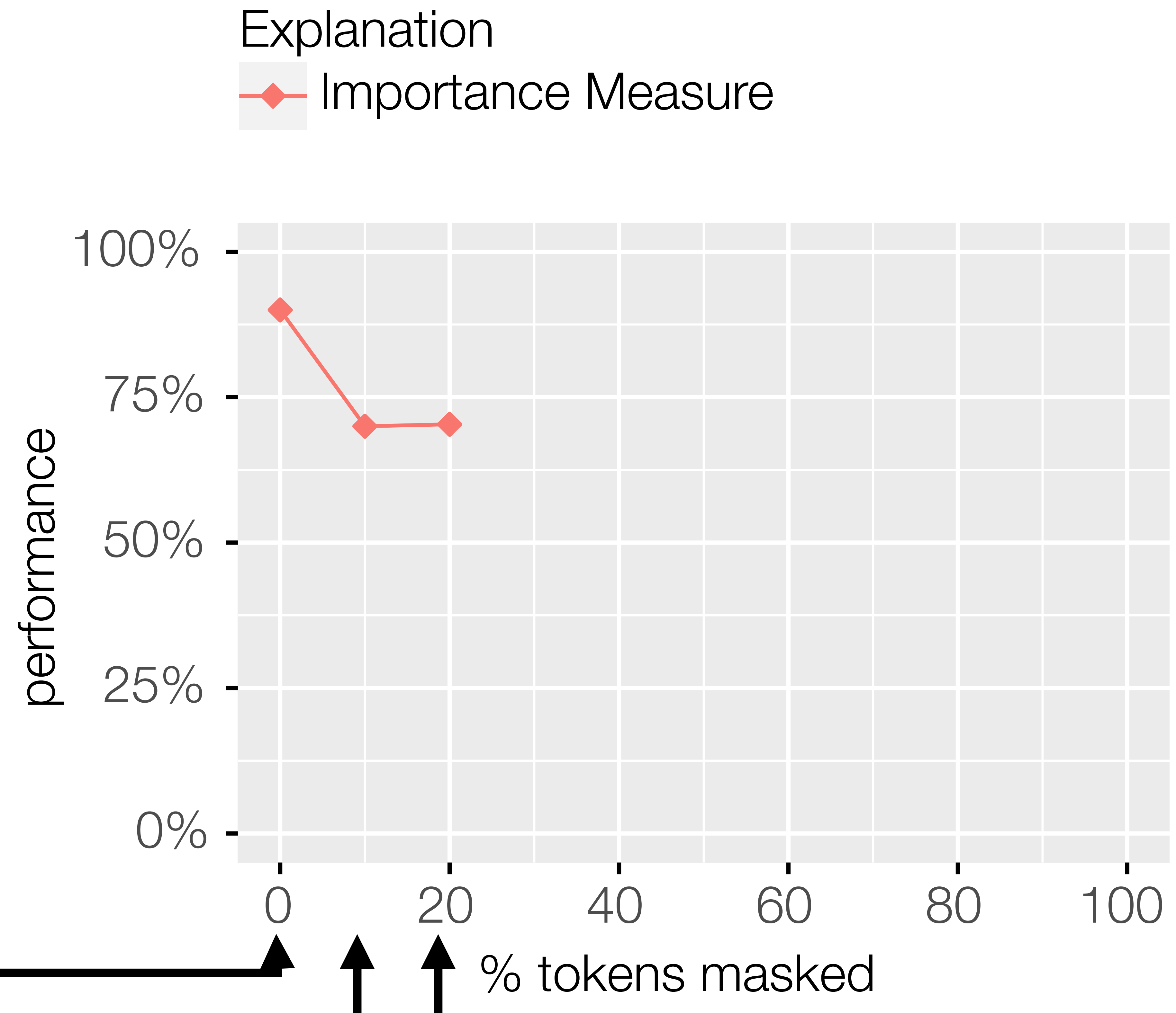
0%	The	movie	was	great	.	I	really	liked	it	.
10%	The	movie	was	*[M]*	.	I	really	liked	it	.

ROAR

Compute importance measure

Repeat this:

1. Mask 10% more of the dataset
2. Retrain the model
3. Measure the performance



0% The movie was great . I really liked it .

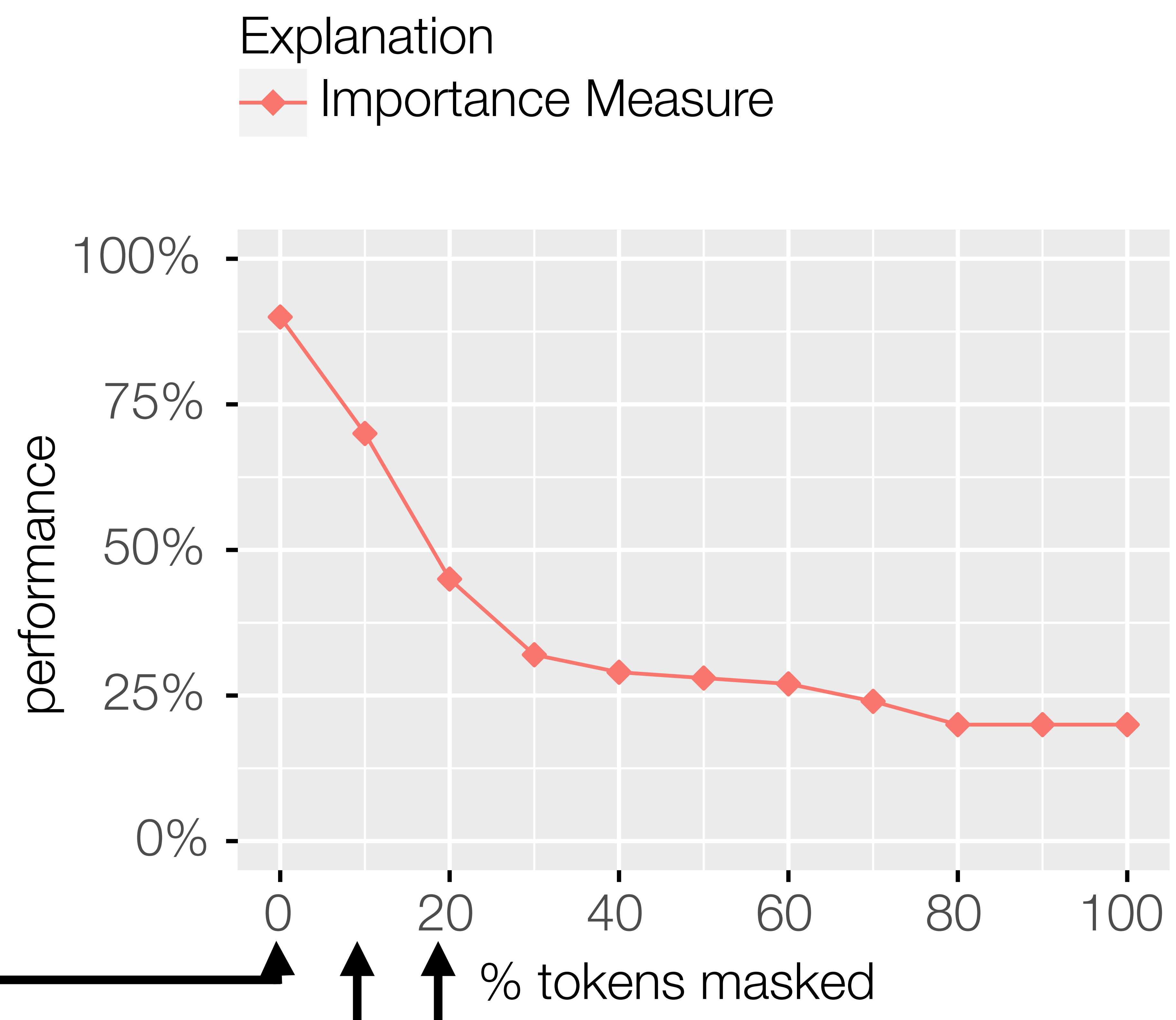
10% The movie was *[M]* . I really liked it .

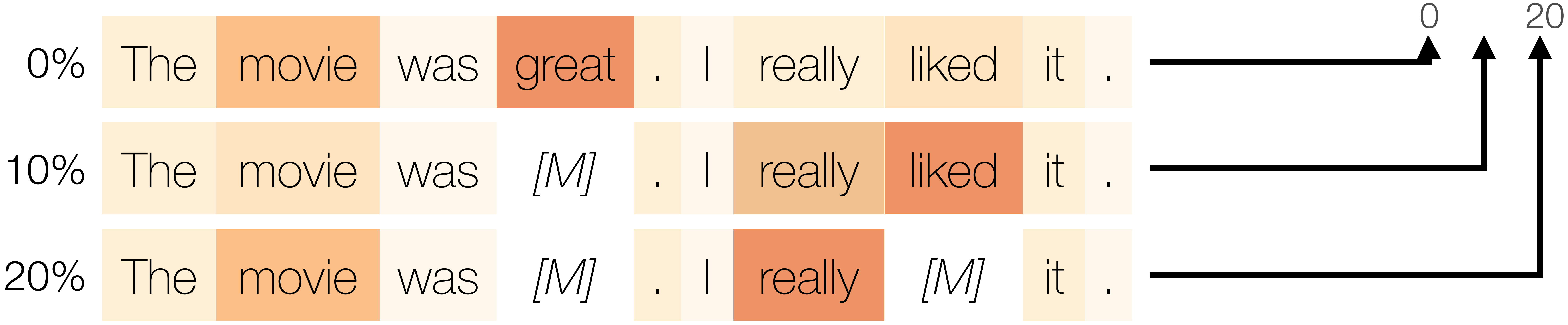
20% The *[M]* was *[M]* . I really *liked* it .

Recursive ROAR

Repeat this:

1. Compute importance measure
2. Mask 10% more of the dataset
3. Retrain the model
4. Measure the performance

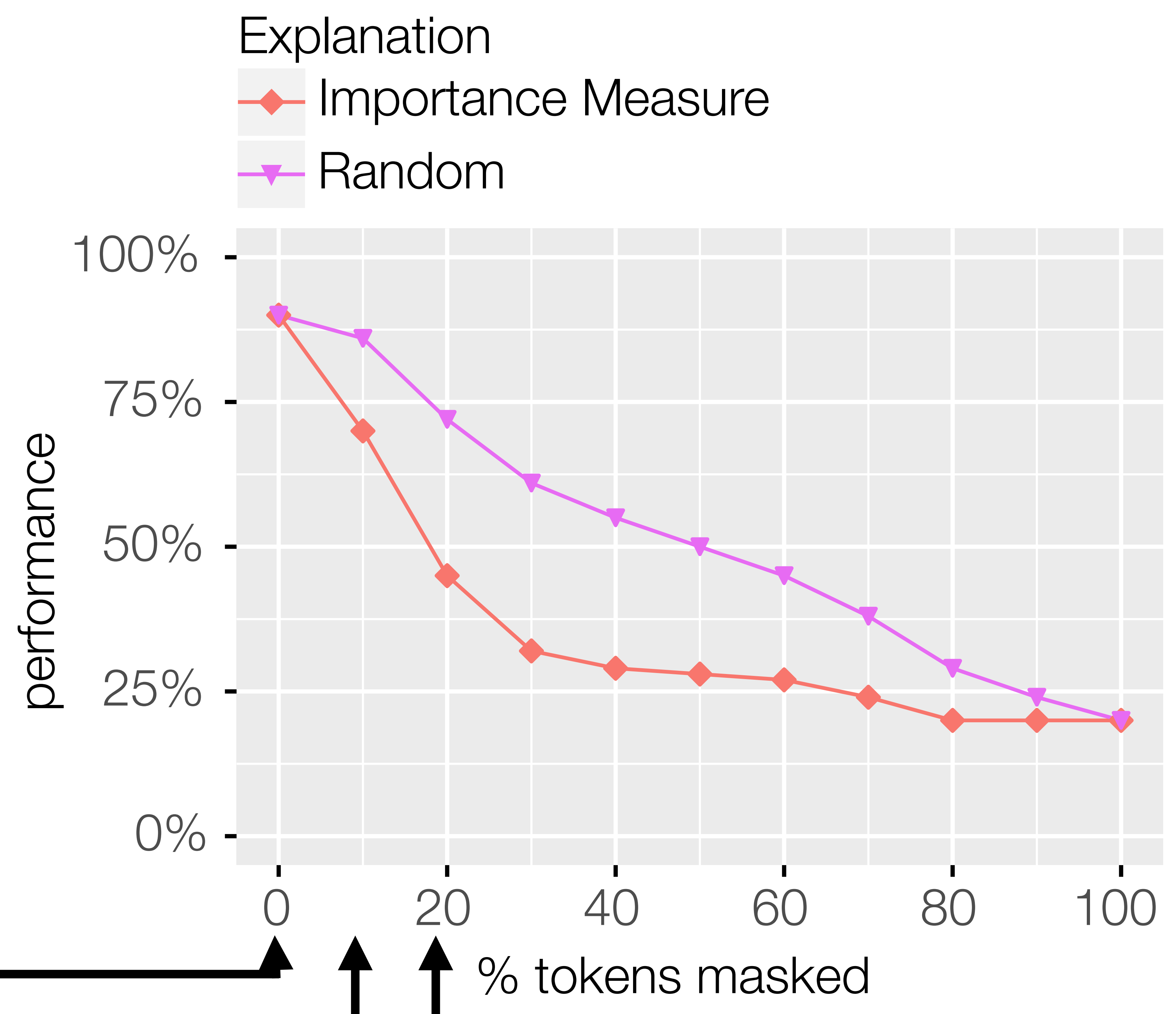




Recursive ROAR

Repeat this:

1. Compute importance measure
2. Mask 10% more of the dataset
3. Retrain the model
4. Measure the performance

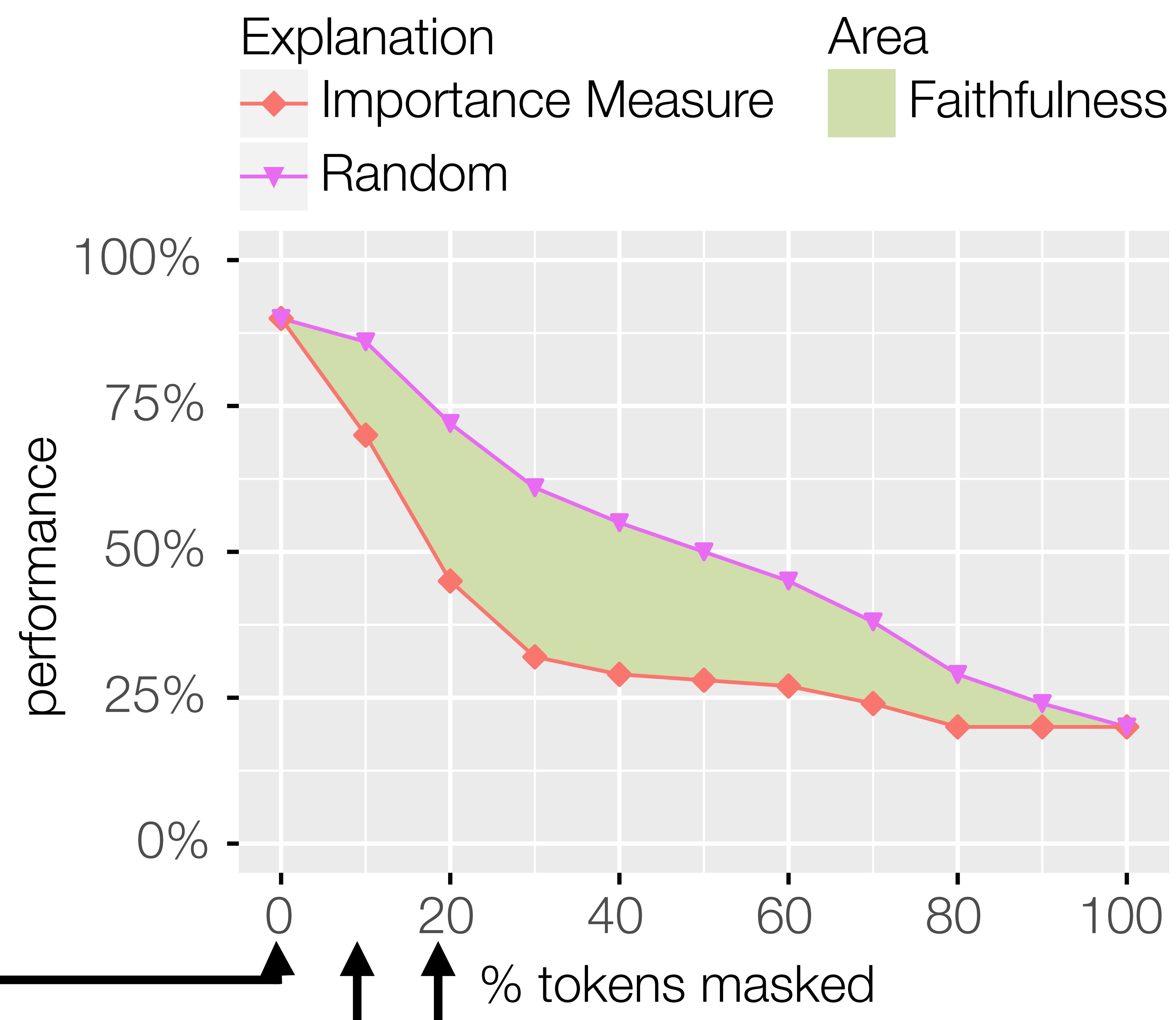


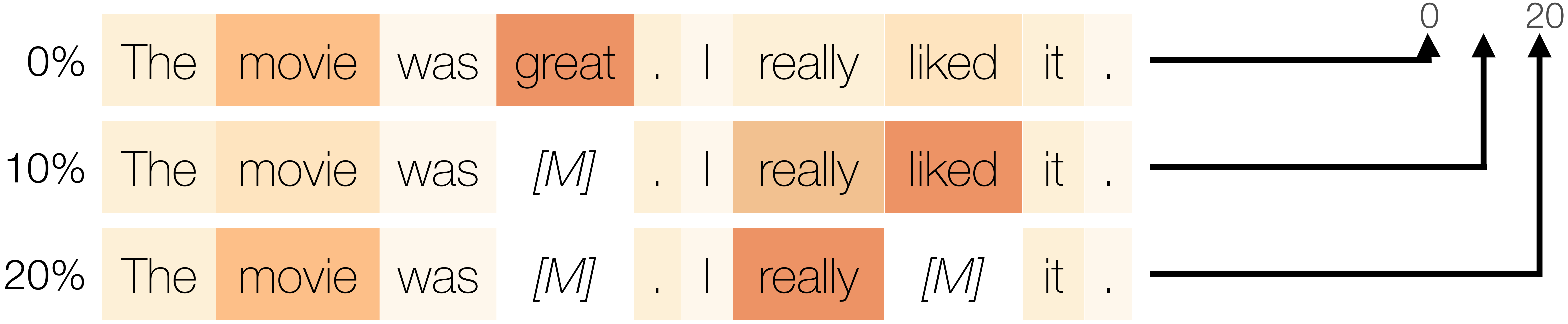
0%	The	movie	was	great	.	I	really	liked	it	.
10%	The	movie	was	[M]	.	I	really	liked	it	.
20%	The	movie	was	[M]	.	I	really	[M]	it	.

Recursive ROAR

Repeat this:

1. Compute importance measure
2. Mask 10% more of the dataset
3. Retrain the model
4. Measure the performance

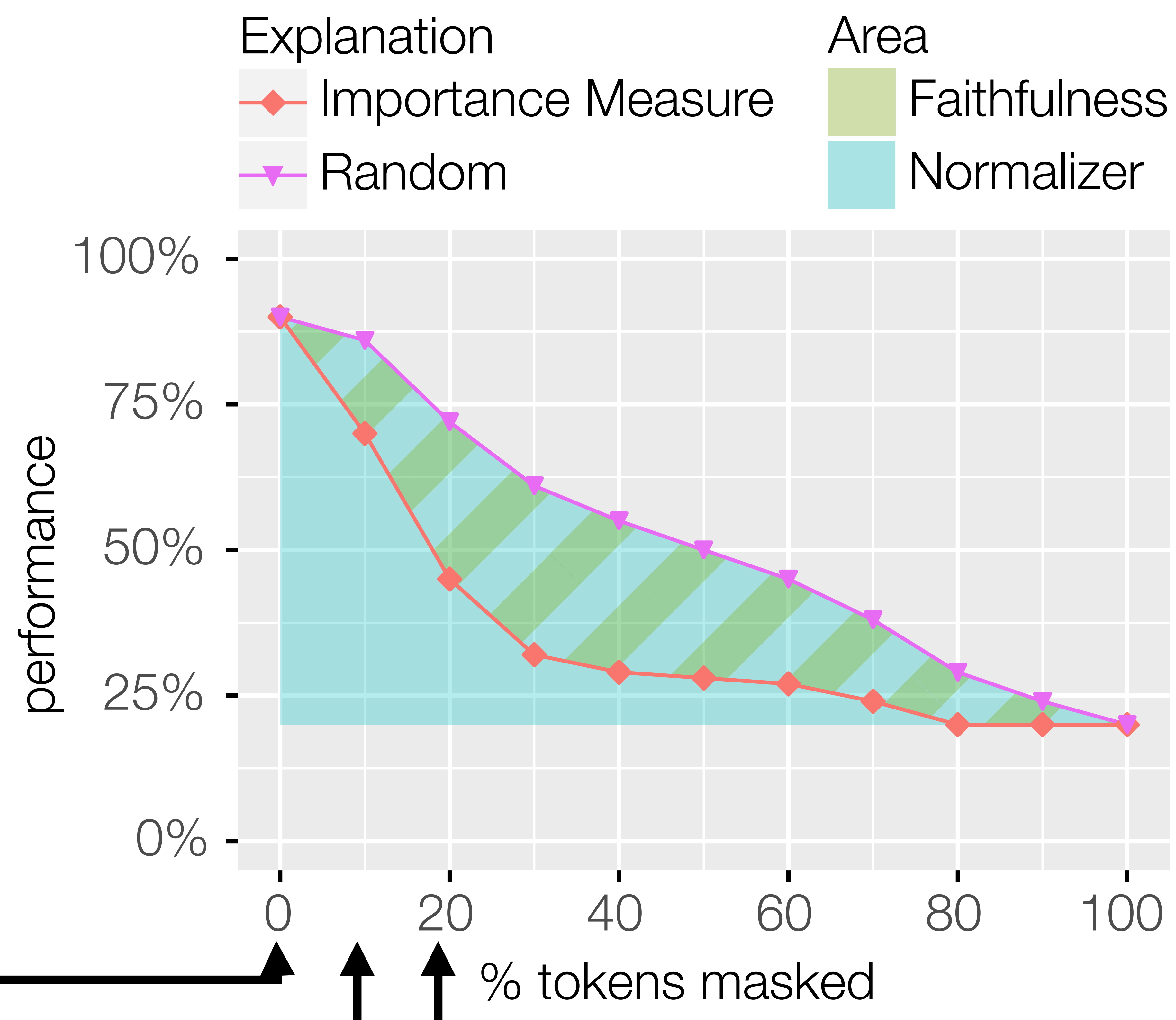


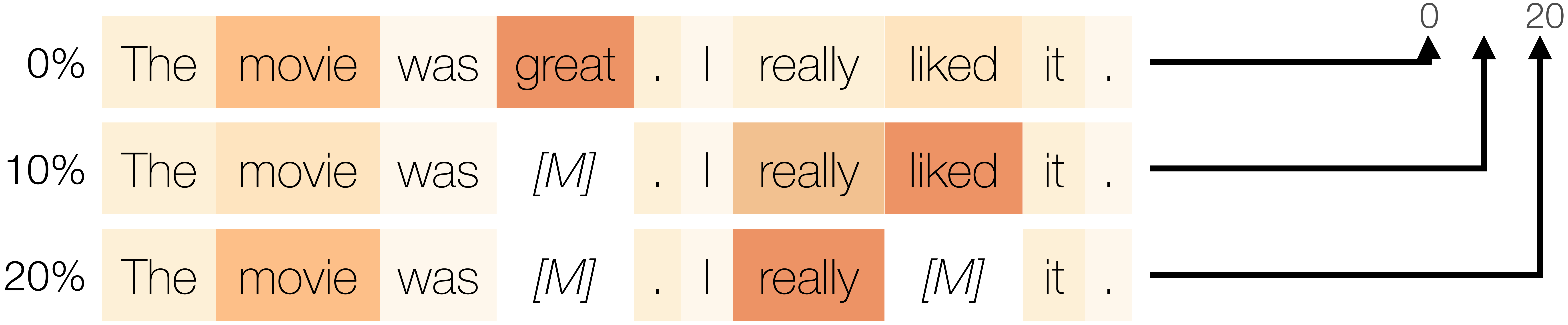


Recursive ROAR

Repeat this:

1. Compute importance measure
2. Mask 10% more of the dataset
3. Retrain the model
4. Measure the performance

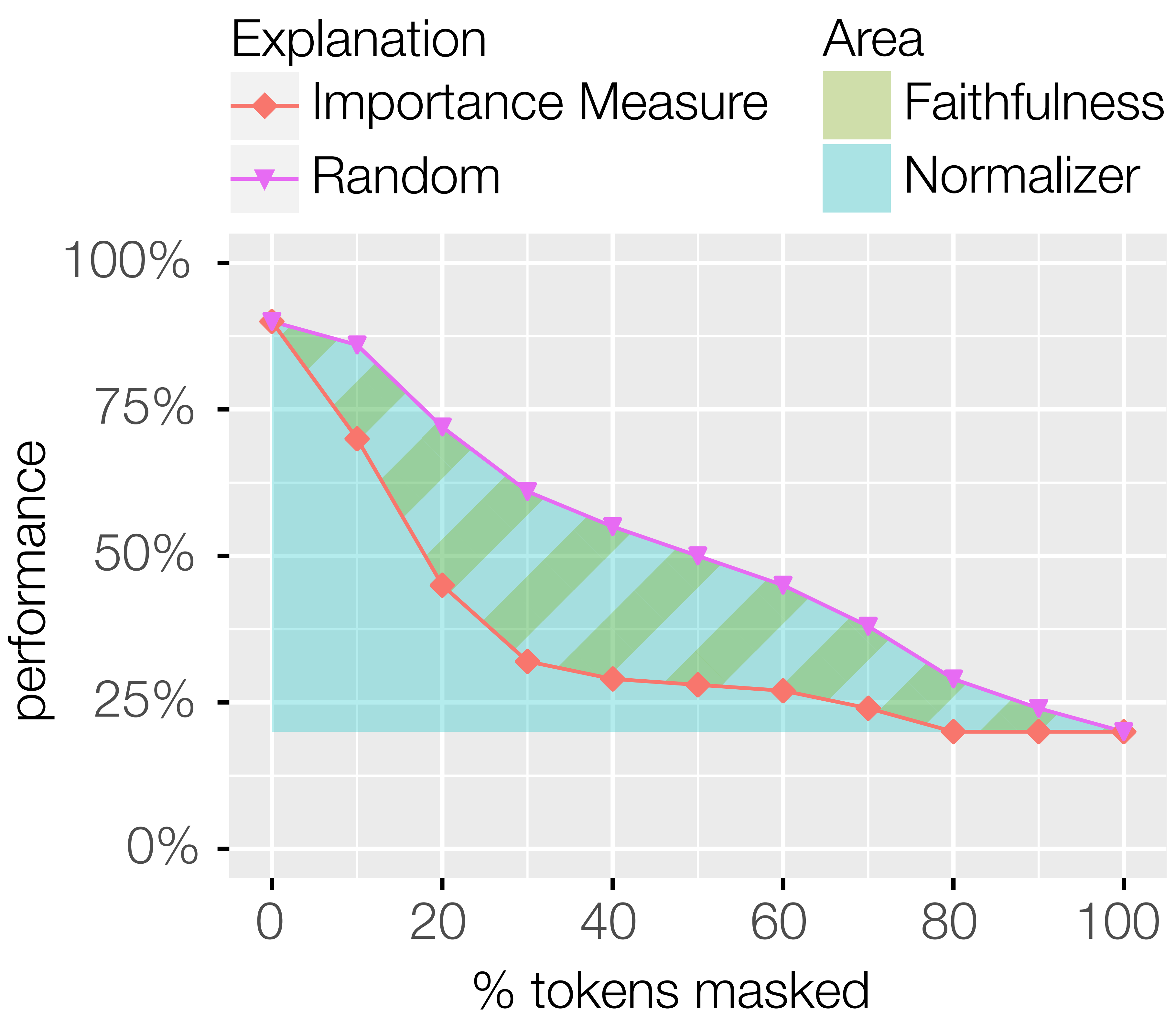




Model and task-dependent faithfulness

	LSTM	RoBERTa
bAbI-1	59.1%	48.2%
bAbI-2	34.6%	42.0%
bAbI-3	25.9%	**-27.9%**
Anemia	4.9%	12.5%
Diabetes	11.4%	26.1%
SST	37.8%	32.9%
SNLI	**-13.9%**	56.7%
IMDB	32.5%	35.1%

Absolute Integrated Gradient



Same conclusion in: Bastings, J., et al. "Will You Find These Shortcuts?" A Protocol for Evaluating the Faithfulness of Input Salience Methods for Text Classification. EMNLP 2022

Limitations

- Computationally expensive:
  - Retrain the model 10 times
  - Importance measure on training dataset
  - For each: explanation, model, and dataset

Limitations

- Computationally expensive:
  - Retrain the model 10 times
  - Importance measure on training dataset
  - For each: explanation, model, and dataset
- Does not measure on the deployed model

Limitations

All because of retraining

- Computationally expensive:
  - Retrain the model 10 times
  - Importance measure on training dataset
  - For each: explanation, model, and dataset
- Does not measure on the deployed model
- Leaks the classification target

What if we had a model that supported masking from the beginning?

Faithfulness Measurable Masked Language Models

Andreas Madsen [1 2] **Siva Reddy** [1 3 4] **Sarath Chandar** [1 2 5]

ICML 2024

Spotlight

Abstract

A common approach to explaining NLP models is to use importance measures that express which tokens are important for a prediction. Unfortunately, such explanations are often wrong despite being persuasive. Therefore, it is essential to measure their faithfulness. One such metric is if tokens are truly important, then masking them should result in worse model performance. However, token masking introduces out-of-distribution issues, and existing solutions that address this are computationally expensive and employ proxy models. Furthermore, other metrics are very limited in scope. This work proposes an inherently faithfulness measurable model that addresses these challenges. This is achieved using a novel fine-tuning method that incorporates masking, such that masking tokens become in-distribution by design. This differs from existing approaches, which are completely model-agnostic but are inapplicable in practice. We demonstrate the generality of our approach by applying it to 16 different datasets and validate it using statistical in-distribution tests. The faithfulness is then measured with 9 different importance measures. Because masking is in-distribution, importance

1. Introduction

As machine learning models are increasingly being deployed, the demand for interpretability to ensure safe operation increases (Doshi-Velez & Kim, 2017). In NLP, importance measures such as attention or integrated gradient are a popular way of explaining which input tokens are important for making a prediction (Bhatt et al., 2019). These explanations are not only used directly to explain models but are also used in other explanations such as contrastive (Yin & Neubig, 2022), counterfactuals (Ross et al., 2021), and adversarial explanations (Ebrahimi et al., 2018).

Unfortunately, importance measures (IMs) are often found to provide false explanations despite being persuasive (Jain & Wallace, 2019; Hooker et al., 2019). For example, a given IMs might not be better at revealing important tokens than pointing at random tokens (Madsen et al., 2022a). This presents a risk, as false but persuasive explanations can lead to unsupported confidence in a model. Therefore, it's important to measure faithfulness. Jacovi & Goldberg (2020) defines faithfulness as: "how accurately it (explanation) reflects the true reasoning process of the model". In this work, we propose a methodology that enables existing models to support measuring faithfulness by design.

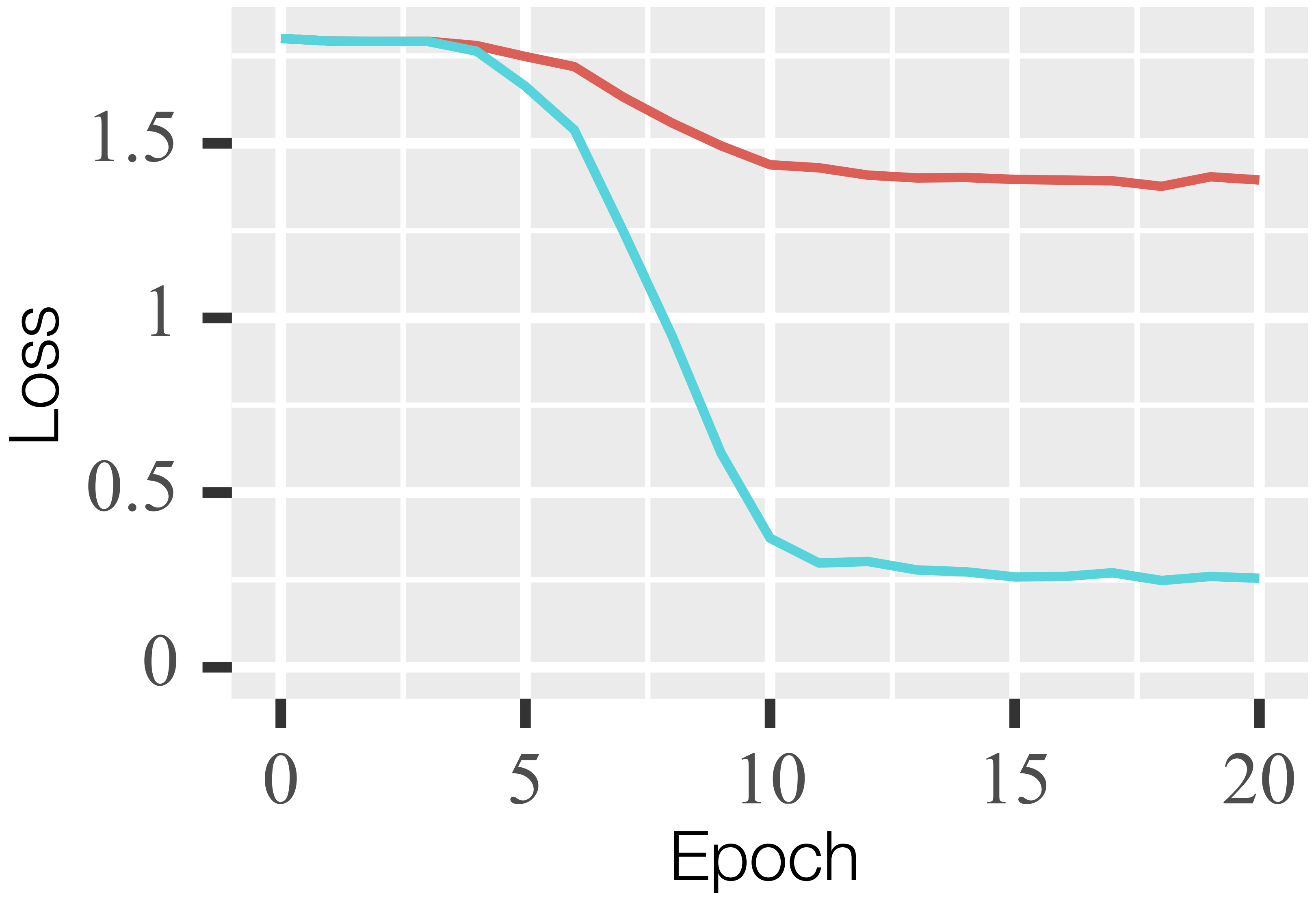
Measuring faithfulness is challenging, as there is generally no known ground-truth for the correct explanation. Instead, faithfulness metrics have to use proxies. One such proxy

Masked Language Models

- Pre-trained with 12% masking (RoBERTa)
- Catastrophic forgetting when fine-tuning

Masked fine-tuning

$$\mathscr{L}\left(X_{1:B}, y_{1:B}\right) = \tilde{\mathscr{L}}\left(X_{1:\frac{B}{2}}, y_{1:\frac{B}{2}}\right) + \tilde{\mathscr{L}}\left(mask\left(X_{\frac{B}{2}:B}\right), y_{\frac{B}{2}:B}\right)$$



0%	The move was great	no masking
0%	Is this acting	
40%	[M] new [M] of comedy	uniform masking
60%	[M] [M] they [M] had	

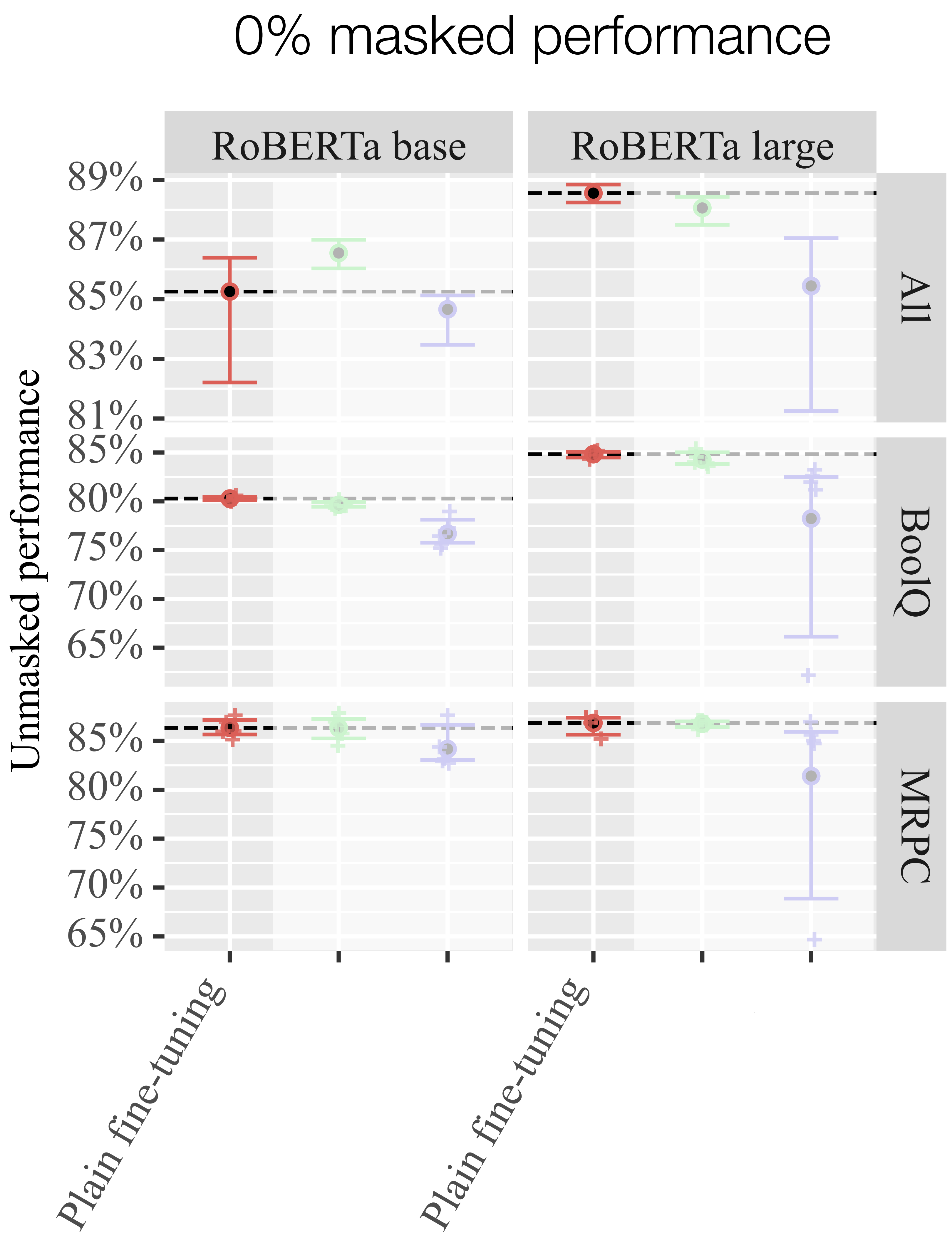
Uniform masking:

In half of the mini-batch.

For each training observation:

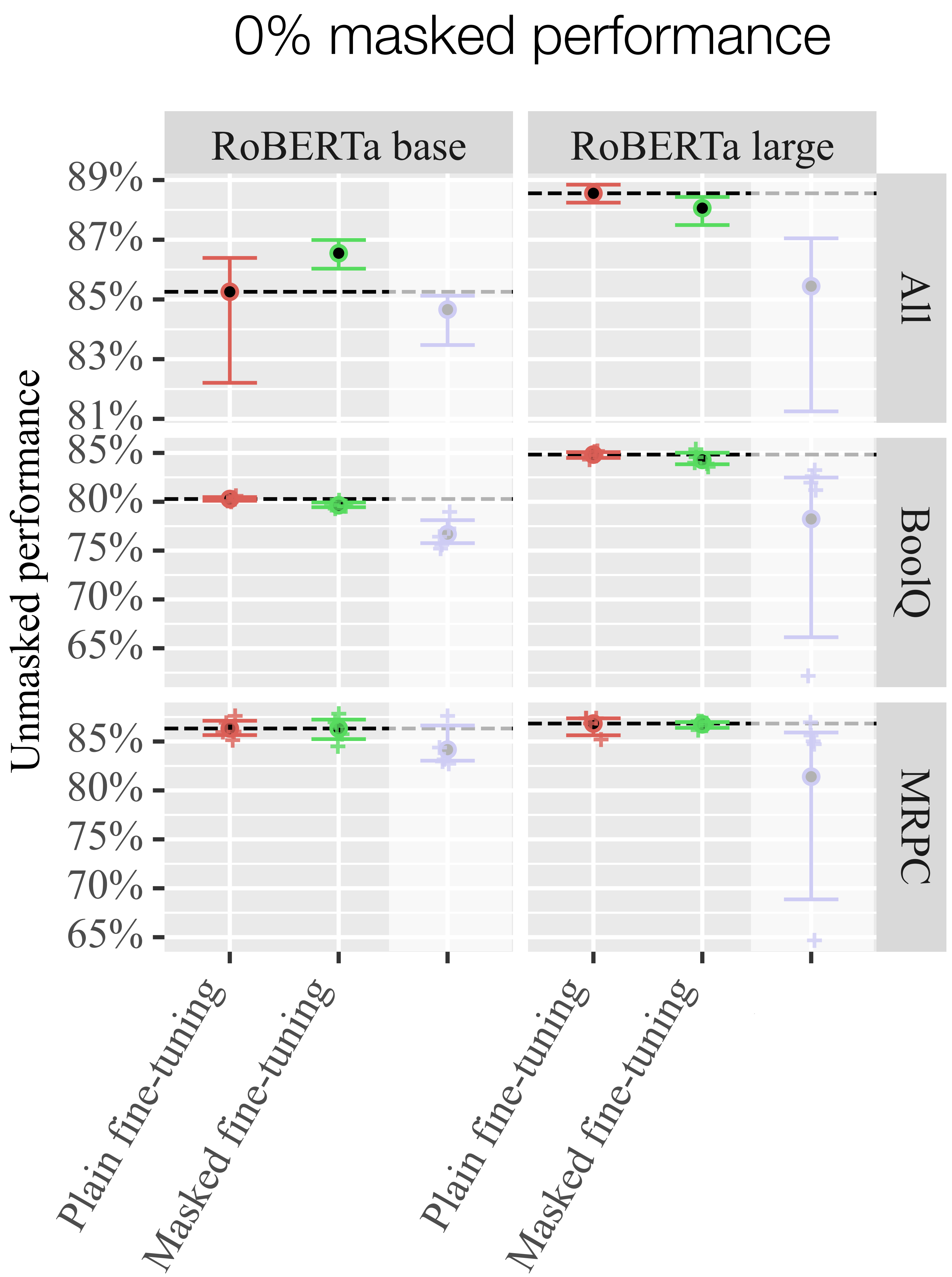
1. Sample a masking ratio between 0% and 100%.
2. Mask random ratio% tokens in an observation.

No performance issues



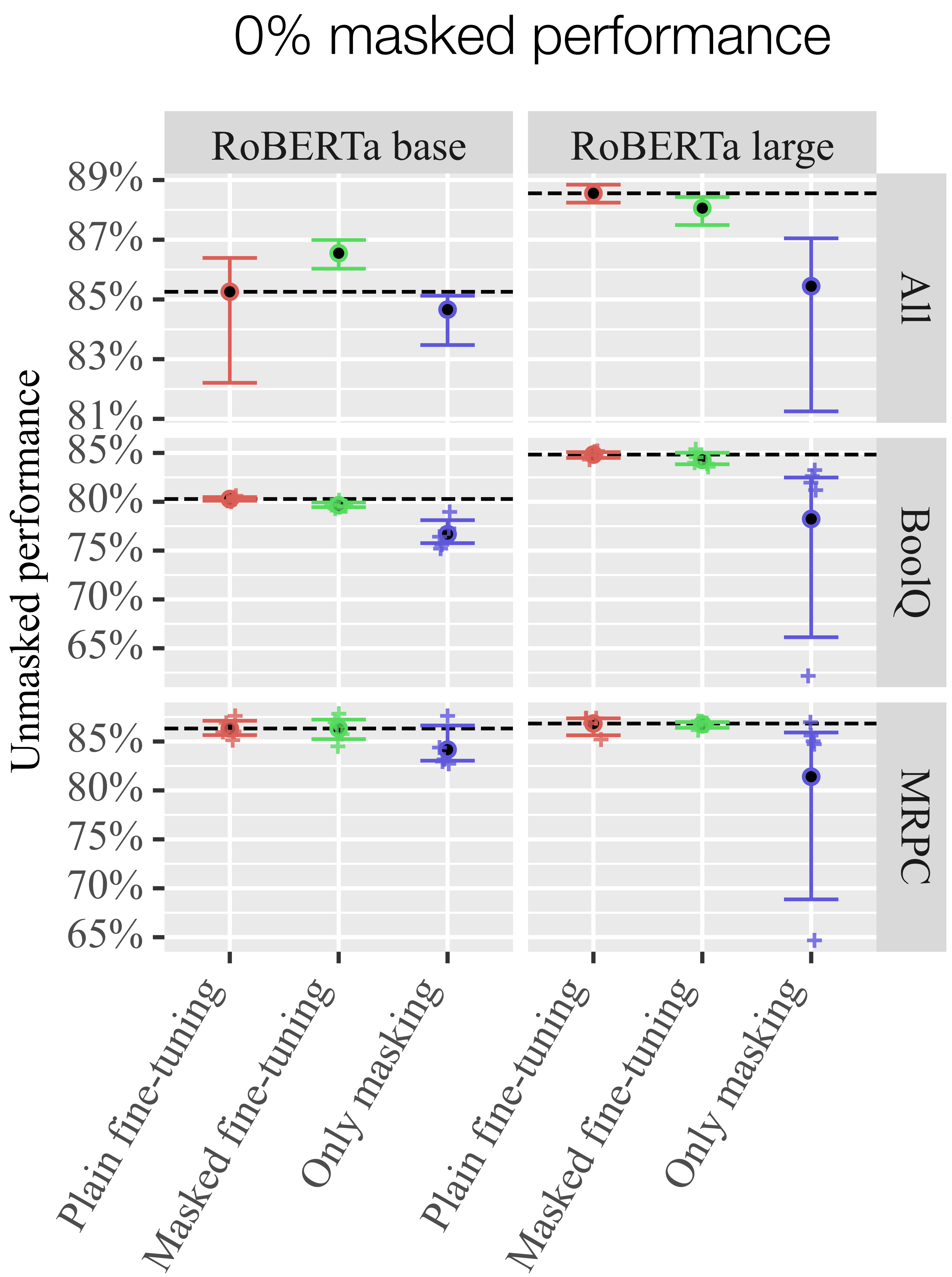
- Default hyperparameters.
- 95% confidence interval of the mean, 5 seeds.

No performance issues



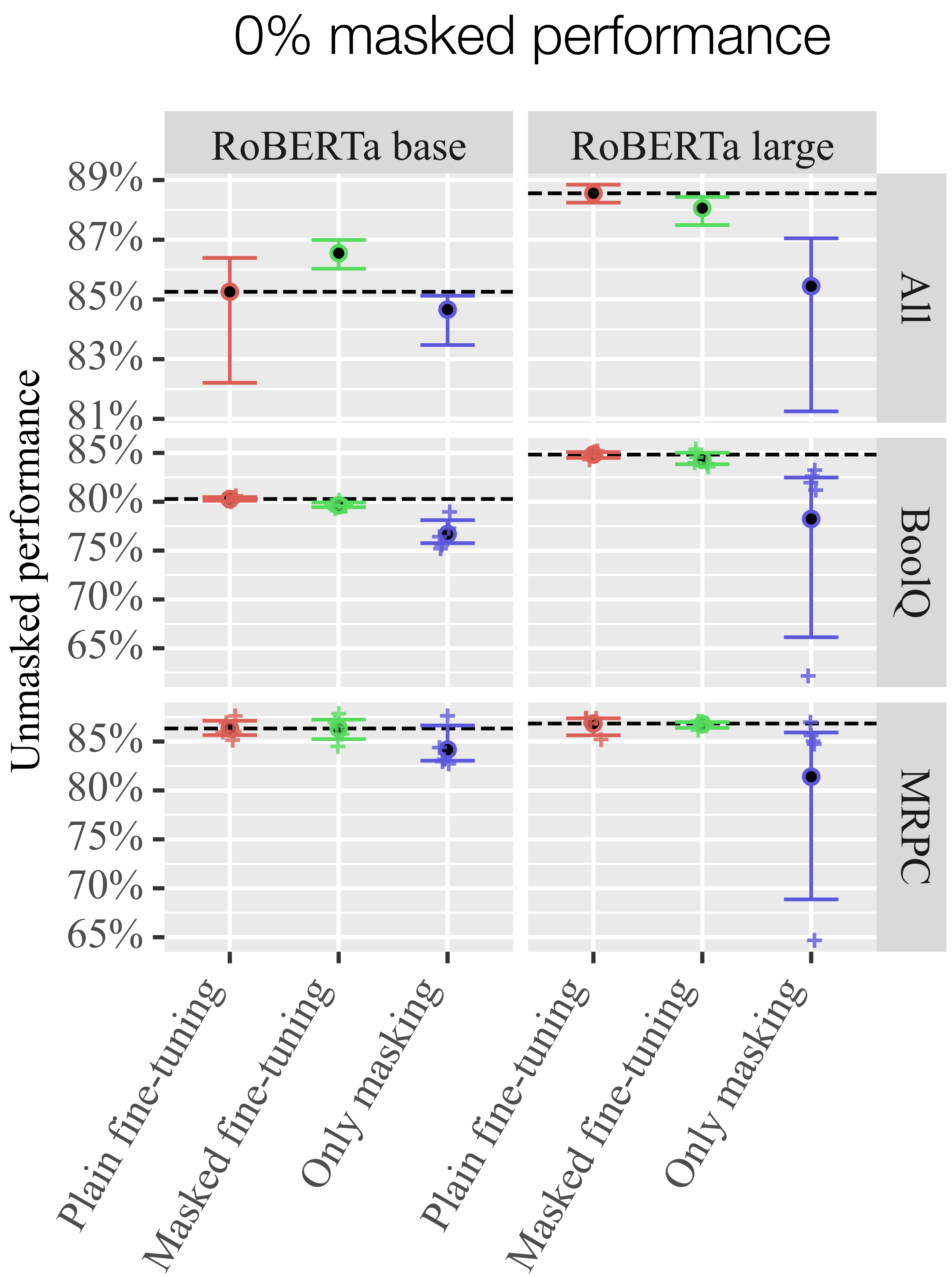
- Default hyperparameters.
- 95% confidence interval of the mean, 5 seeds.

No performance issues



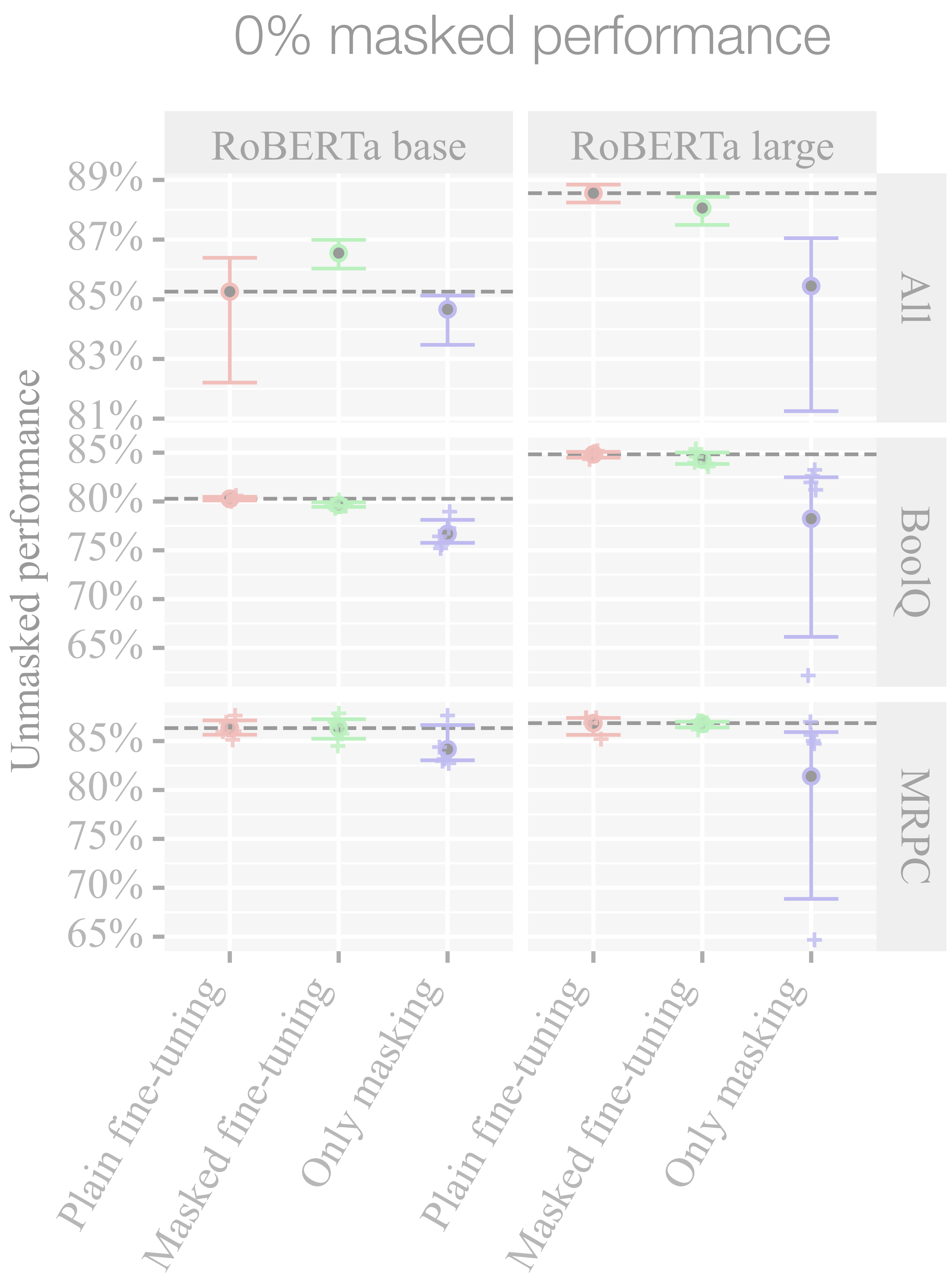
- Default hyperparameters.
- 95% confidence interval of the mean, 5 seeds.

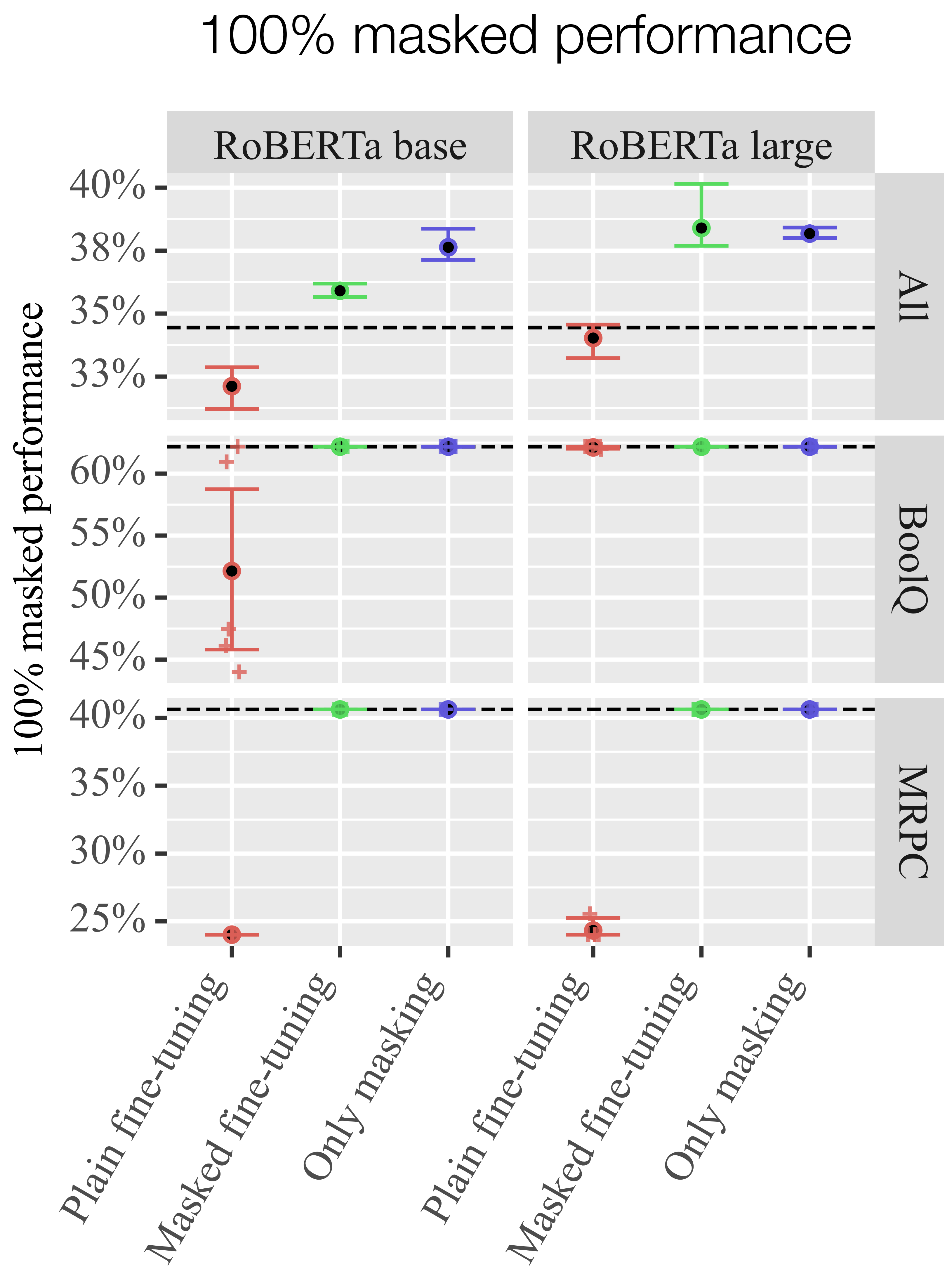
No performance issues



Type	Dataset
NLI	RTE SNLI MNLI CB
Parahrase	MRPC QQP
Sentiment	SST2 IMDB
Diagnosis	Anemia Diabetese
Acceptability	CoLA
QA	BoolQ bAbI-1 bAbI-2 bAbI-3

No performance issues





✅ ? ✅

0% 100%

masking-ratio

In-distribution testing

- Should assume little of the model's internals. For example, do not assume internally normally distributed.
- Should only consider the model, not the input distribution.
- Should provide non-ambiguous metrics.

In-distribution testing

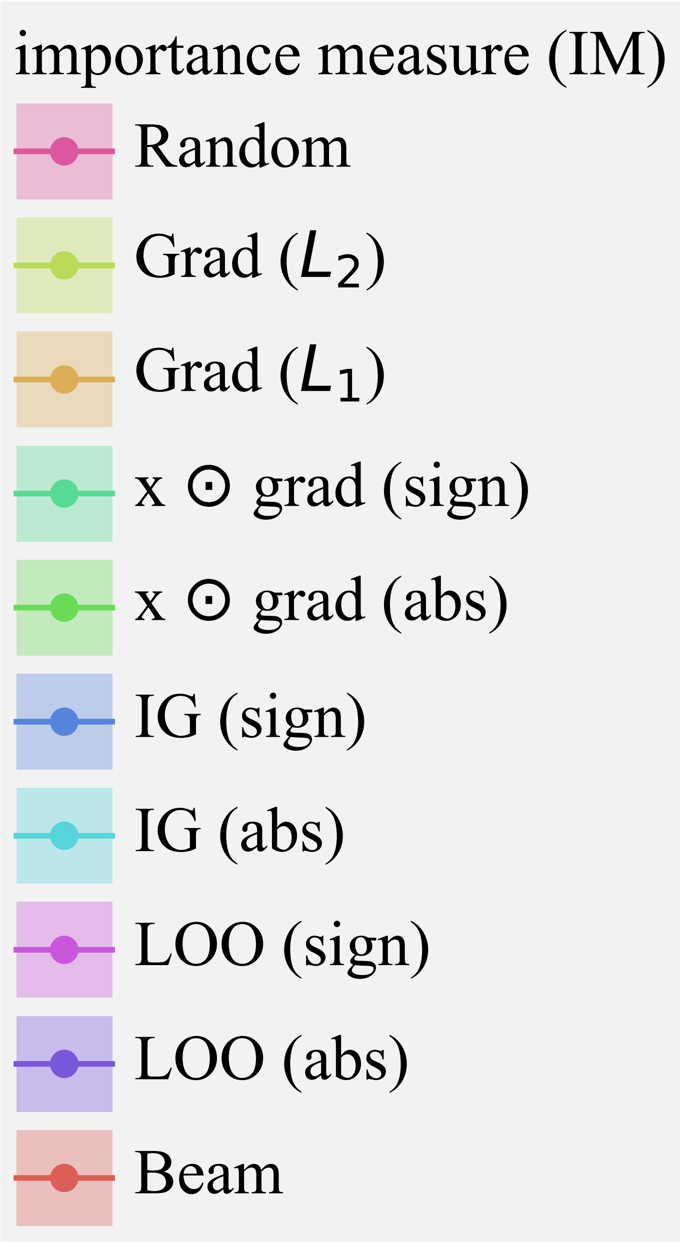
- Should assume little of the model's internals. For example, do not assume internally normally distributed.
- Should only consider the model, not the input distribution.
- Should provide non-ambiguous metrics.

- Use MaSF [1], a non-parametric statistical global in-distribution test.
- Originally made for small scale computer vision, which we adapt to large scale NLP.

[1] Matan, H., Frostig, T., Heller, R., & Soudry, D. A Statistical Framework for Efficient Out of Distribution Detection in Deep Neural Networks. ICLR 2022

In-distribution testing

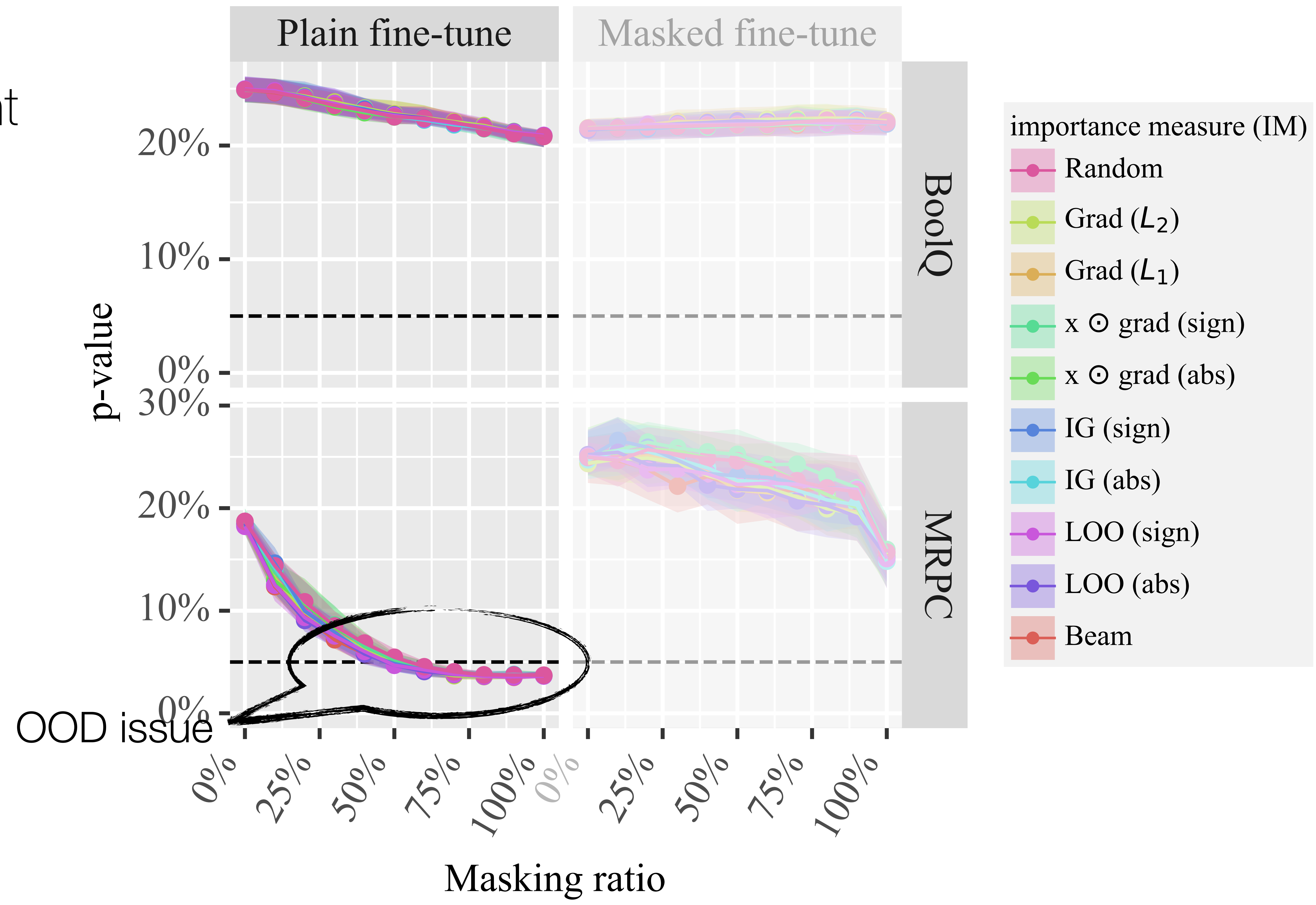
- Because random masking is different form targeted masking,
  each explanation need to be tested.



In-distribution testing

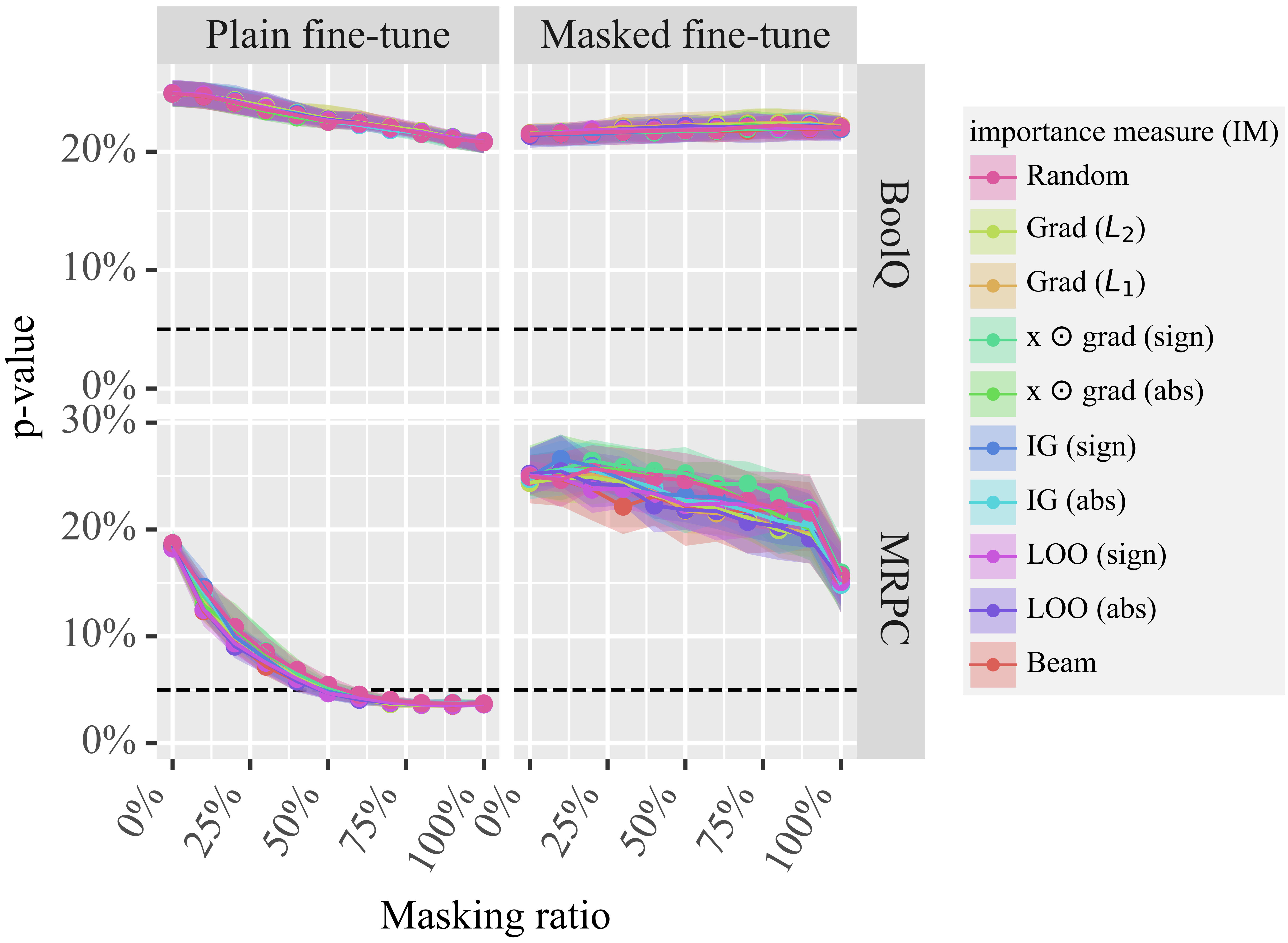
- Because random masking is different form targeted masking,
  each explanation need to be tested.

- Often out-of-distribution issues with plain fine-tuning.

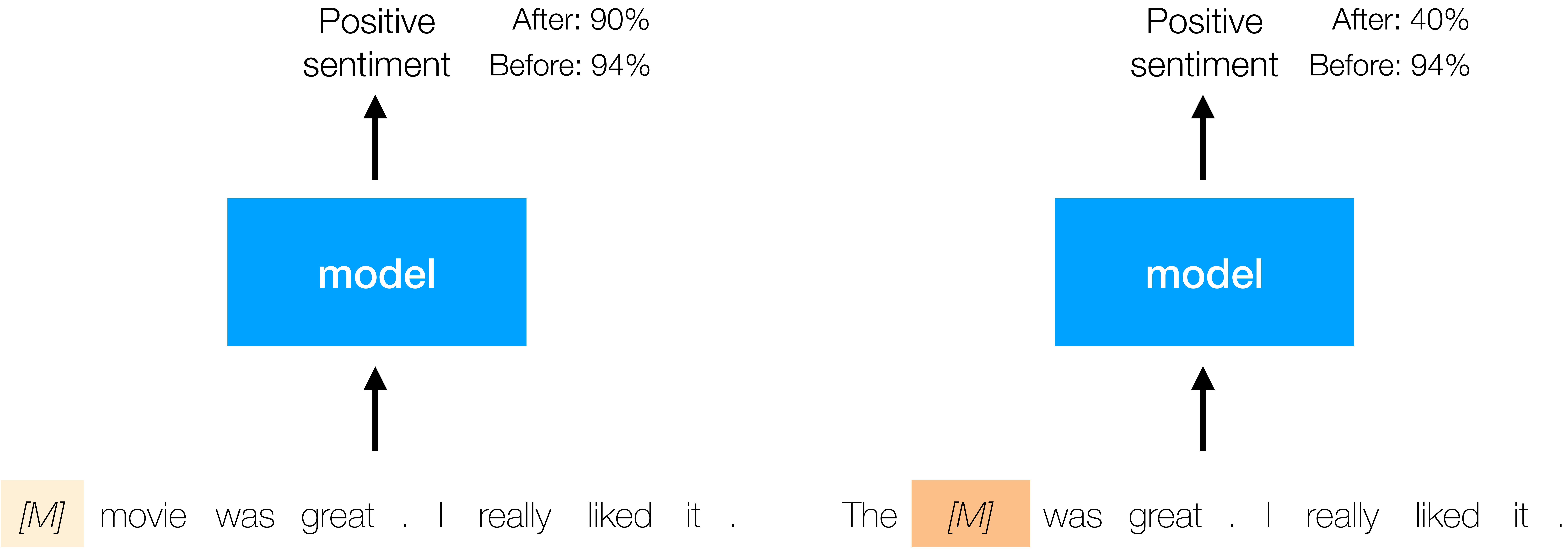


In-distribution testing

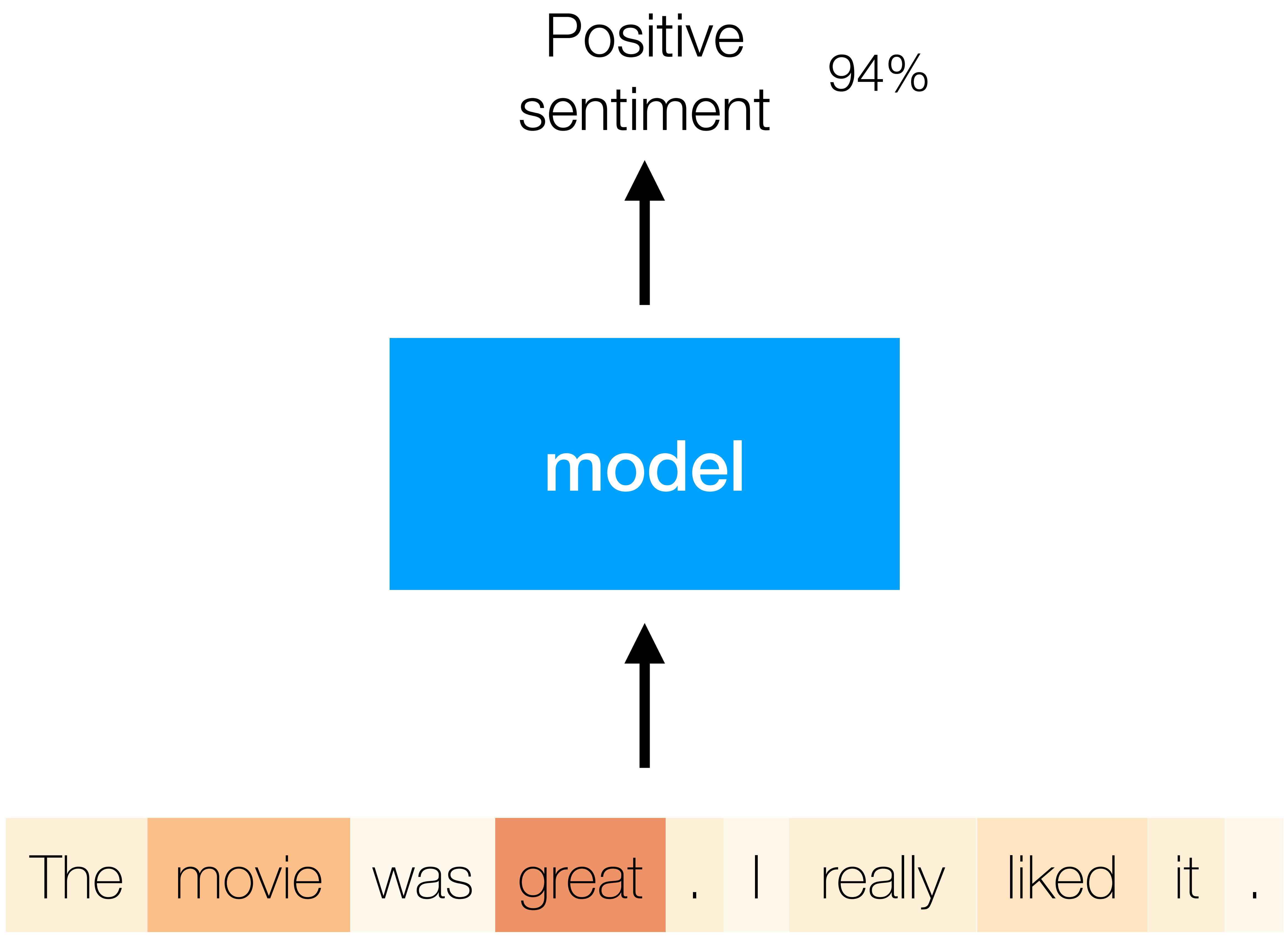
- Because random masking is different form targeted masking,
  each explanation need to be tested.
- Often out-of-distribution issues with plain fine-tuning.
- No out-of-distribution issues with masked fine-tuning.



Occlusion-based



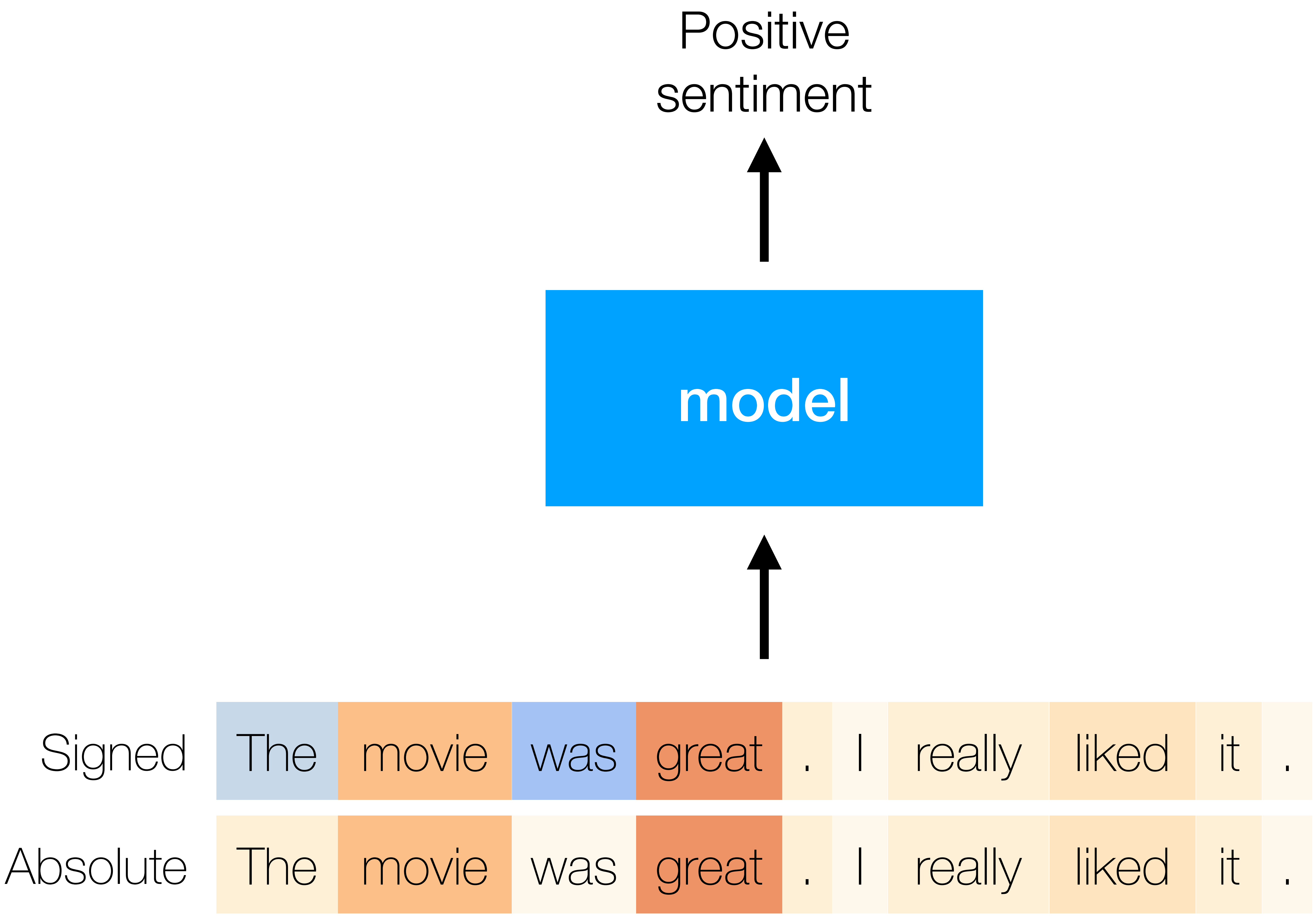
Gradient-based



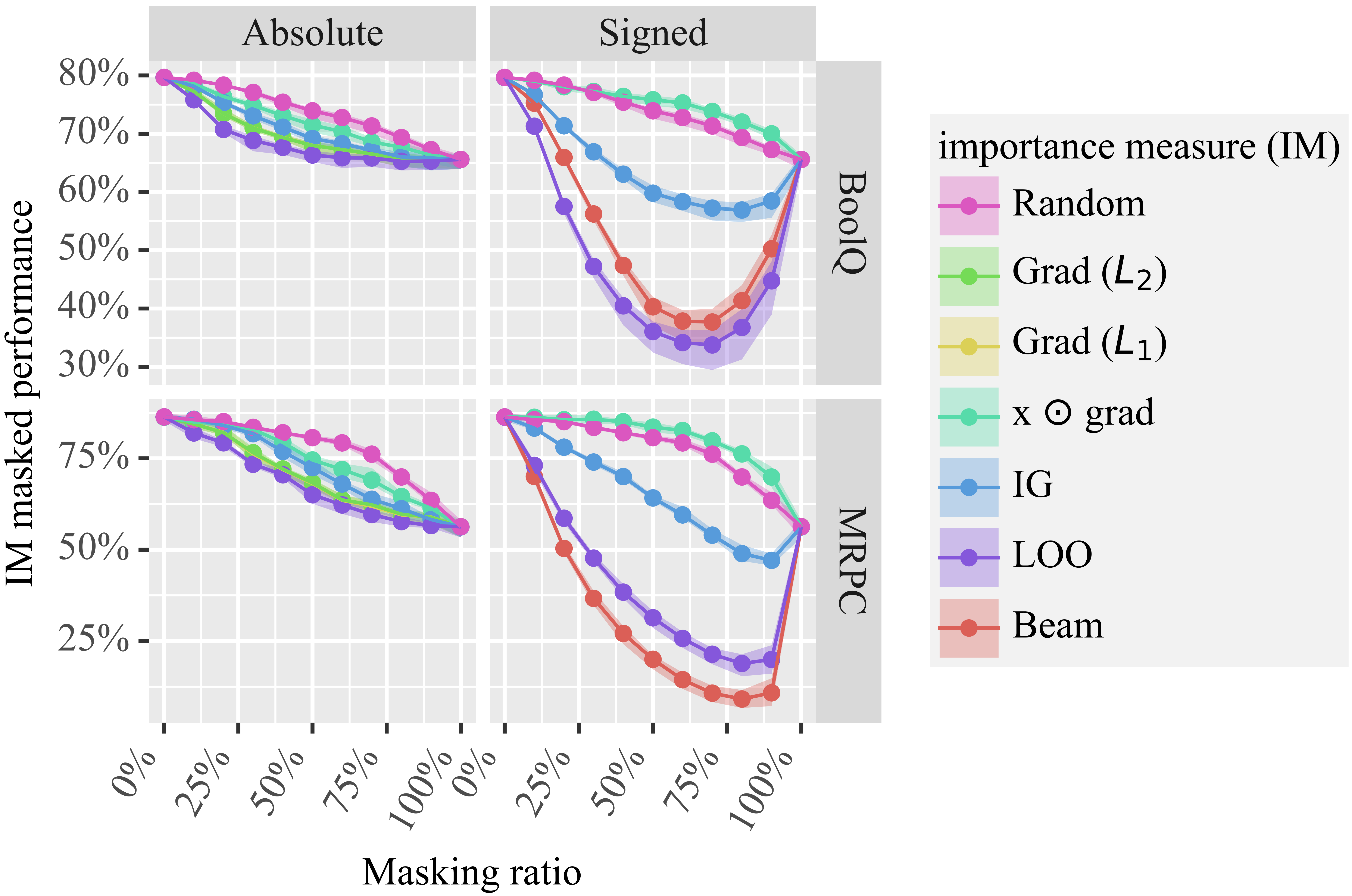
$$e(x_i) = \frac{\partial f(x)_c}{\partial x_i}$$

x

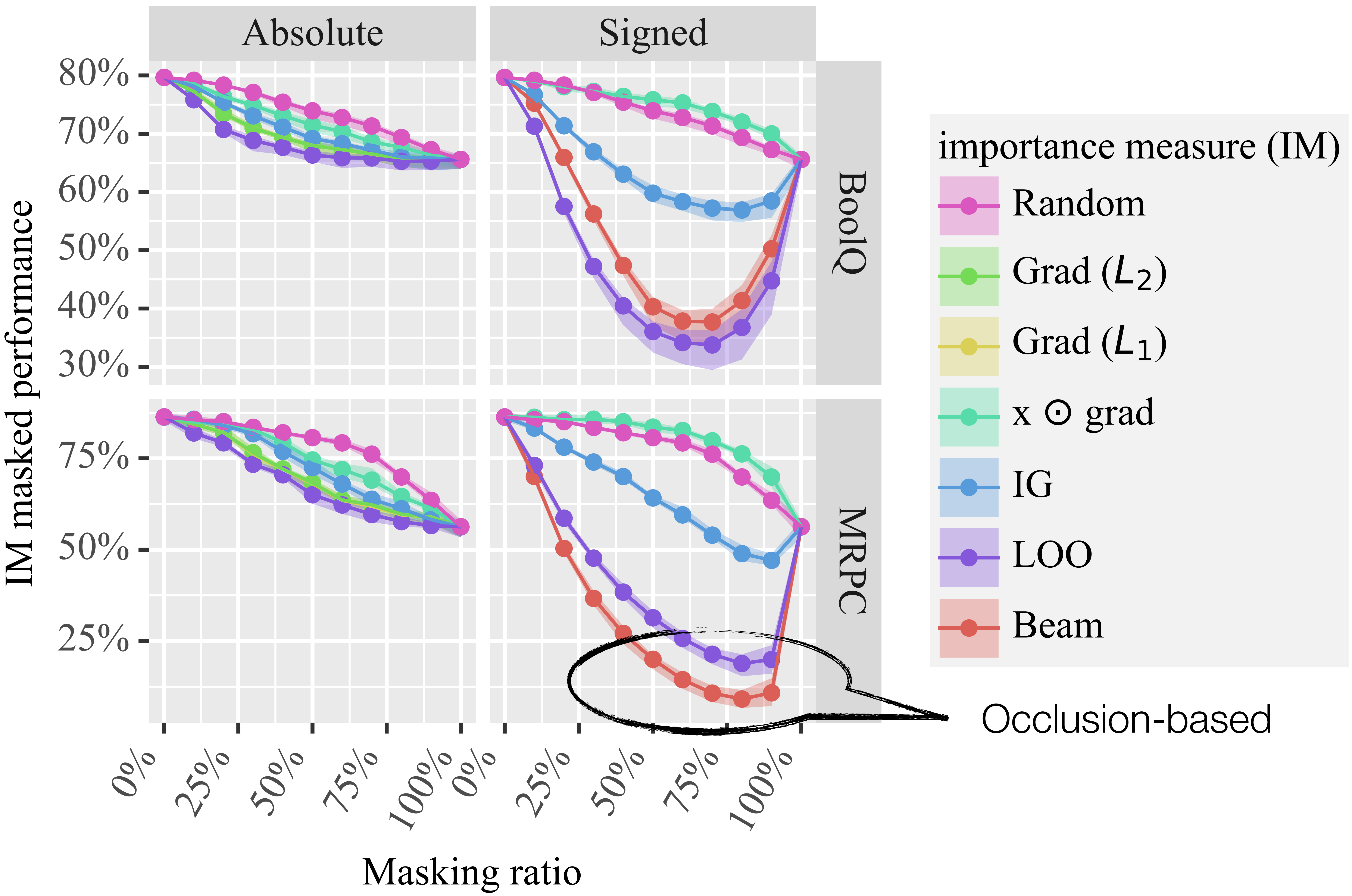
Importance Measures



Faithfulness



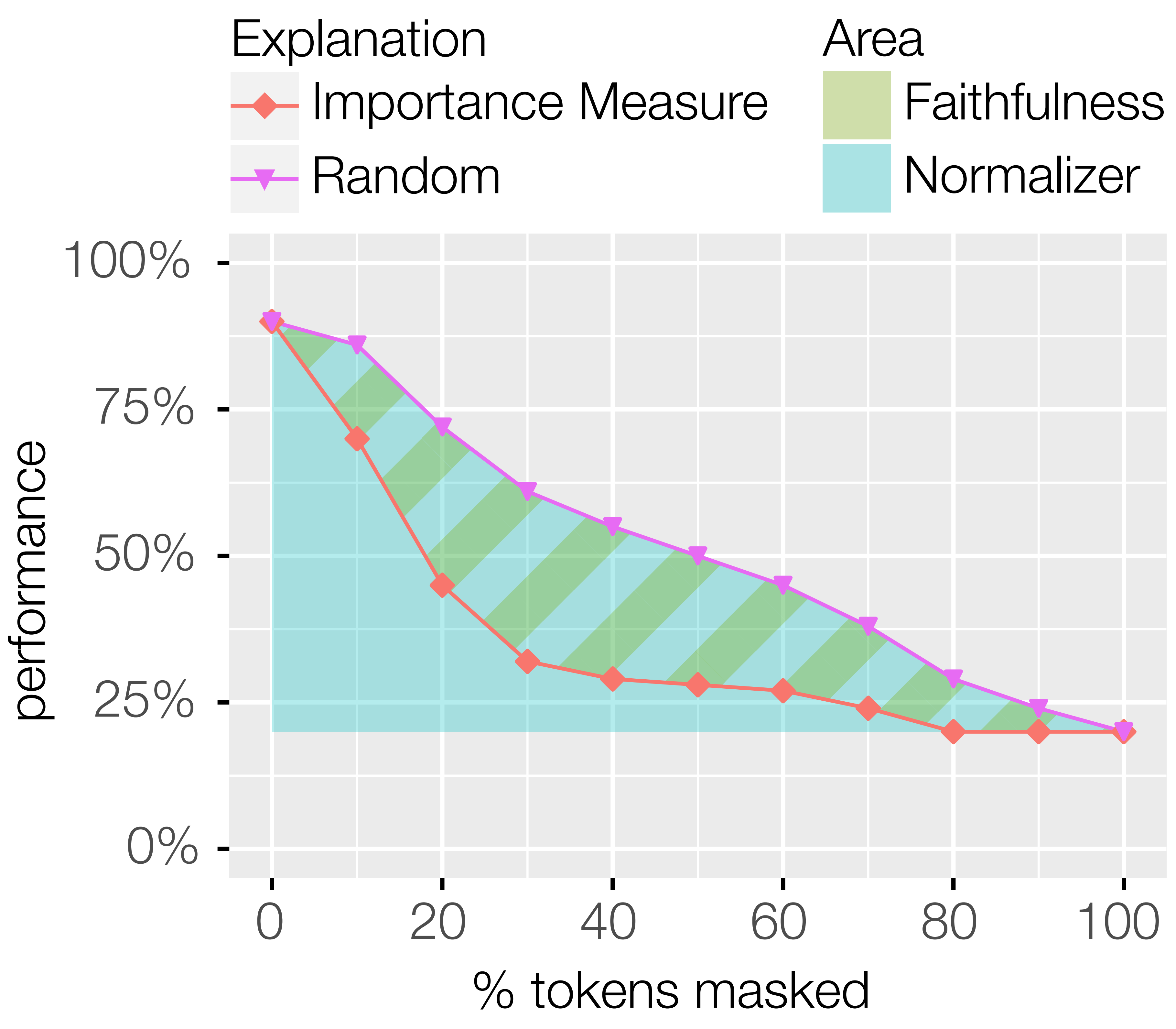
Faithfulness



Comparison

Dataset	IM	FMM	R-ROAR
SST2	Grad (L2)	40.4%	26.1%
	X ⊙ grad (abs)	23.5%	18.6%
	IG (abs)	45.3%	32.9%
bAbI-2	Grad (L2)	96.3%	57.8%
	X ⊙ grad (abs)	92.0%	48.1%
	IG (abs)	98.3%	42.0%

RoBERTa-Base



Higher faithfulness

Dataset	IM	FMM	R-ROAR
SST2	Grad (L2)	40.4%	26.1%
	X ⊙ grad (abs)	23.5%	18.6%
	IG (abs)	45.3%	32.9%
bAbI-2	Grad (L2)	96.3%	57.8%
	X ⊙ grad (abs)	92.0%	48.1%
	IG (abs)	98.3%	42.0%

RoBERTa-Base

John went to the office.
Mary went to the hallway.
John went to the **bathroom**.

Where is John?

Higher faithfulness

Dataset	IM	FMM	R-ROAR
SST2	Grad (L2)	40.4%	26.1%
	X ⊙ grad (abs)	23.5%	18.6%
	IG (abs)	45.3%	32.9%
bAbI-2	Grad (L2)	96.3%	57.8%
	X ⊙ grad (abs)	92.0%	48.1%
	IG (abs)	98.3%	42.0%

RoBERTa-Base

[M] went *[M] [M] [M].*
[M] [M] to *[M] [M]*.
John *[M]* to *[M]* **bathroom**.

- Produces a more robust model, that depends on more relevant signals.
- Faithful explanations then reveals objectively important information.

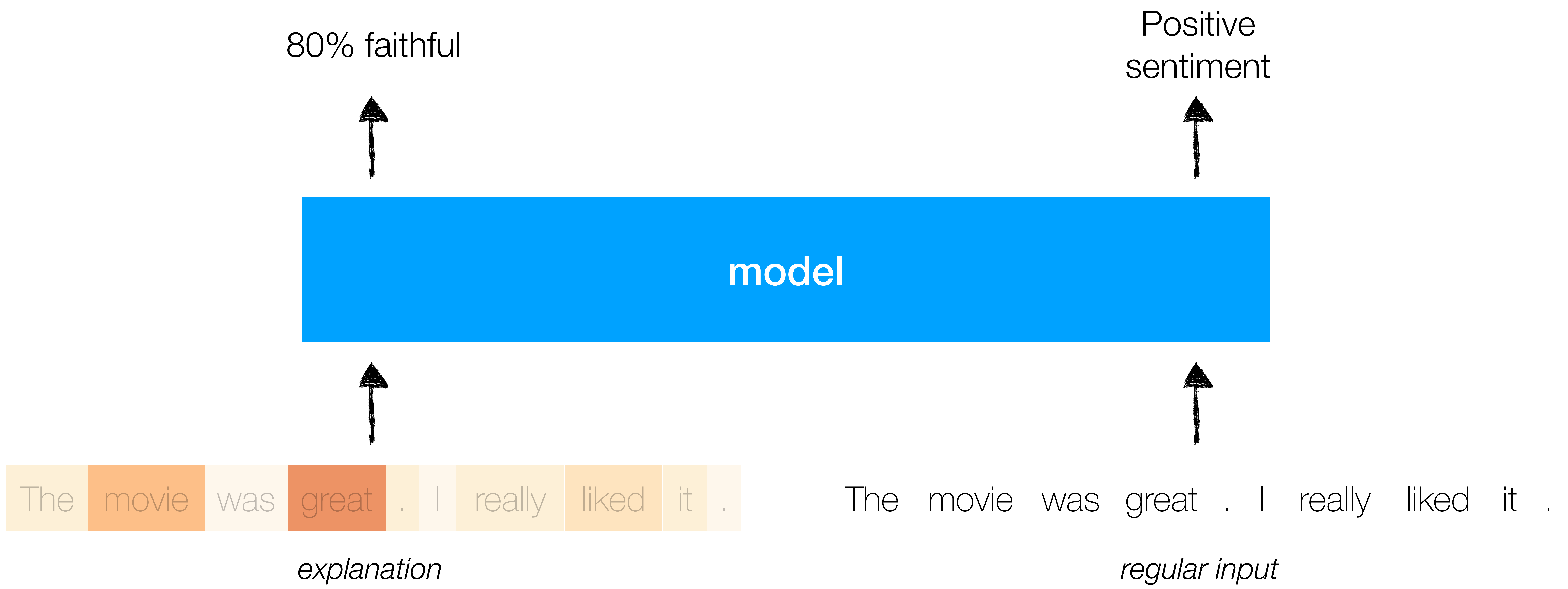
Not model and task-dependent

Dataset	IM	FMM	R-ROAR
bAbI-1	IG (abs)	93.7%	48.2%
bAbI-2		98.3%	42.0%
bAbI-3		100 %	-27.9%
Anemia	IG (abs)	52.1%	12.5%
Diabetes		90.5%	26.1%
SST	IG (abs)	45.3%	32.9%
SNLI		92.3%	56.7%
IMDB		35.4%	35.1%

RoBERTa-Base

- Improvements across all datasets.
- There are now consistently good importance measures, across all 16 datasets.

Faithfulness measurable model



Optimizing for faithfulness

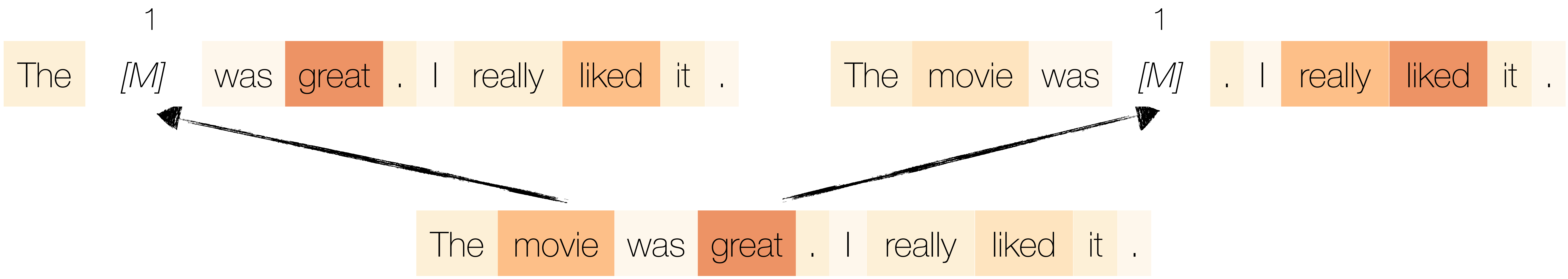
- Building on existing work which uses a beam-search optimizer [1].

- Slightly different faithfulness metric. They use comprehensiveness – sufficiency, we use Recursive ROAR, but same idea.

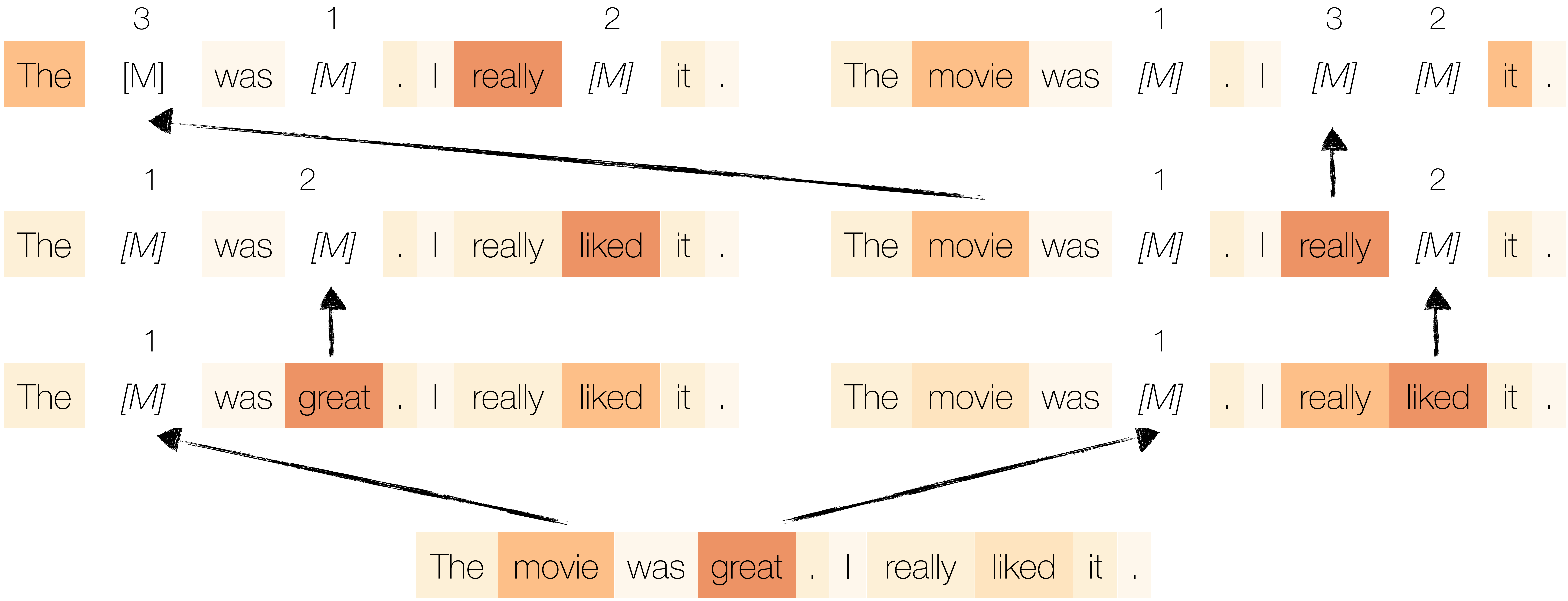
- They do not address the OOD issues caused by masking.

[1] Zhou, Y., & Shah, J. The Solvability of Interpretability Evaluation Metrics. EACL Findings, 2023.

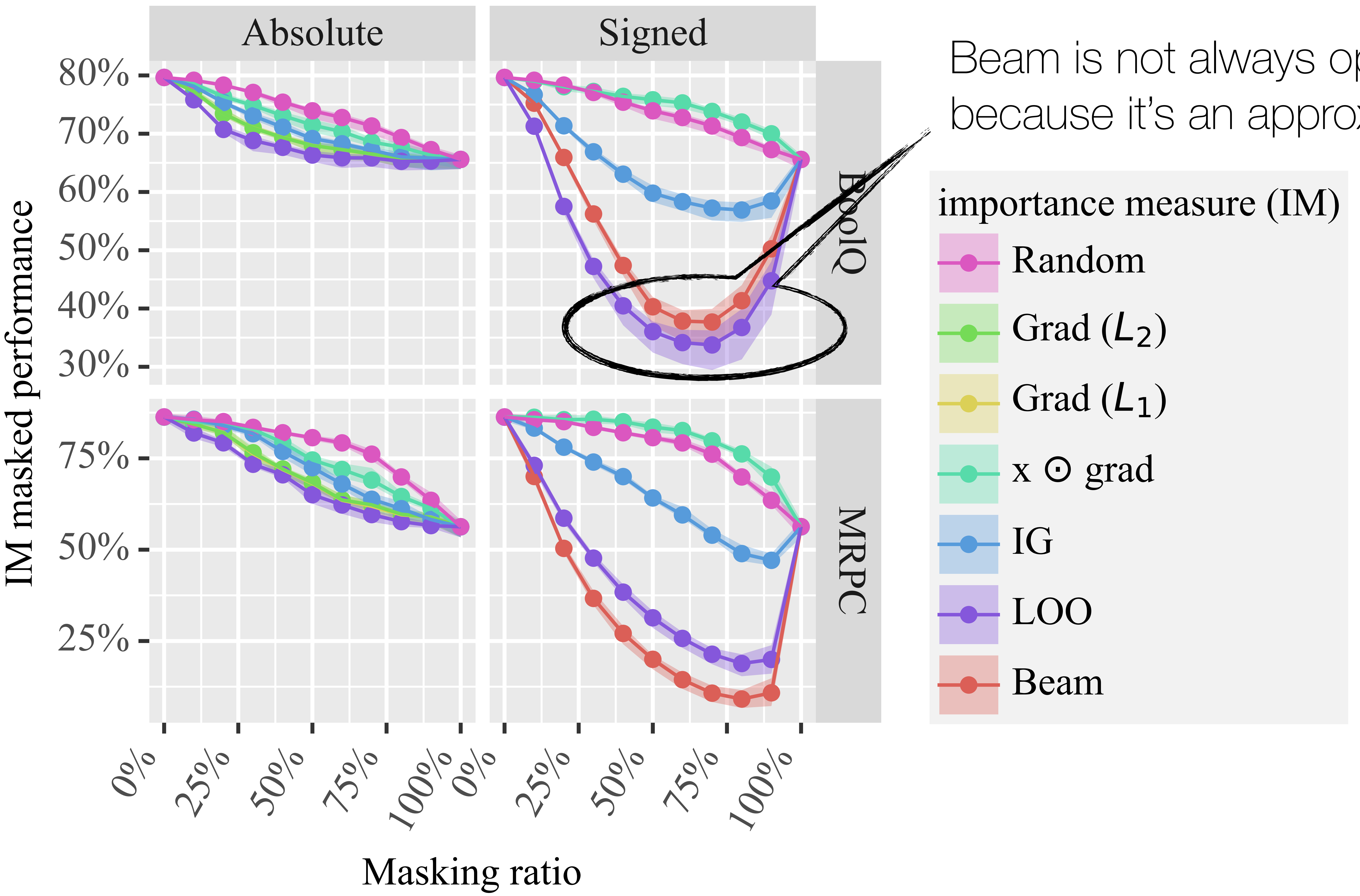
Optimizing for faithfulness



Optimizing for faithfulness



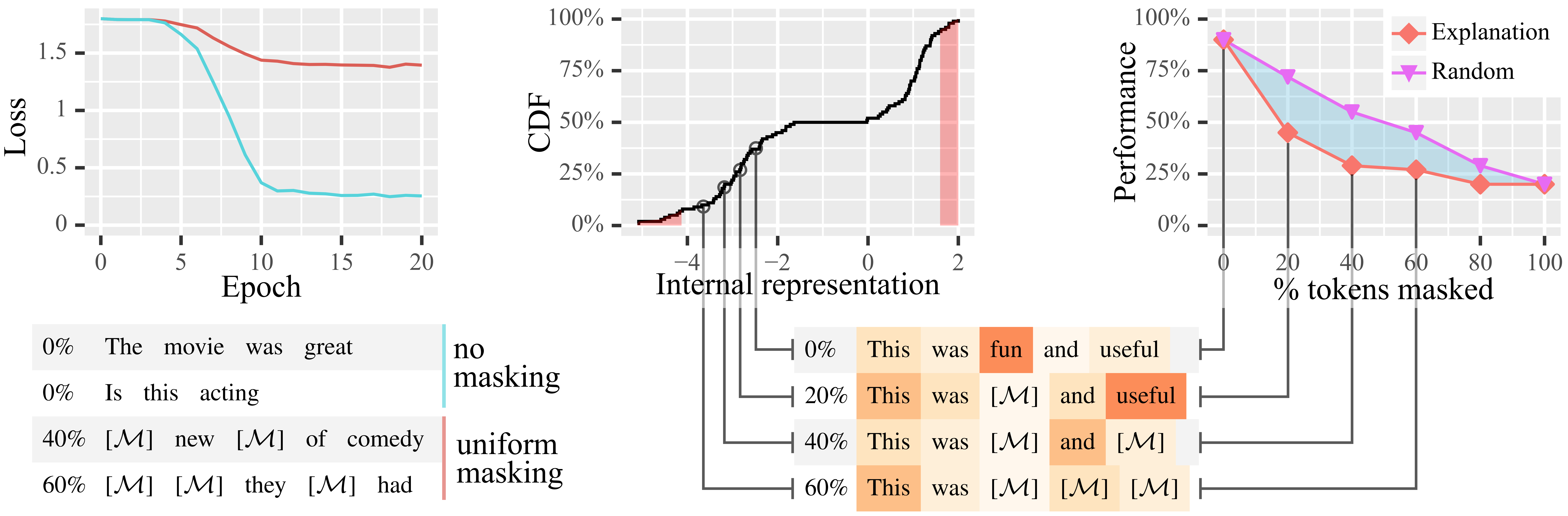
Faithfulness



Beam is not always optimal, because it's an approximation.

Summary

1. Masked fine-tuning 2. In-distribution validation 3. Measure faithfulness



Faithfulness Measurable Models

Only models designed to be explained can be explained.

Black-box models are more general purpose.

less information → more information

post-hoc | intrinsic

lower abstraction ↓ higher abstraction

local explanation	black-box	dataset	gradient	embeddings	white-box	model specific
input features	Occlusion -based § 2.5.2		Gradient -based § 2.5.1			Attenton -based § 2.5.3
adversarial examples	SEA$^{\mathcal{M}}$ § A.1.2		HotFlip § A.1.1			
influential examples		Influence Functions$^{\mathcal{H}}$ § A.2.1 TracIn$^{\mathcal{C}}$ § A.2.3		Representer Pointers† § A.2.2		Prototype Networks
counter-factuals	Polyjuice$^{\mathcal{M},\mathcal{D}}$ § 2.6.1	MiCE$^{\mathcal{M}}$ § 2.6.2				
natural language	predict-then-explain$^{\mathcal{M}}$ § 2.7.2					explain-then-predict$^{\mathcal{M}}$ § 2.7.1
class explanation						
concepts					NIE$^{\mathcal{D}}$ § A.3.1	
global explanation						
vocabulary				Project § A.4.1, Rotate § A.4.2		
ensemble	SP-LIME § A.5.1					
linguistic information	Behavioral Probes$^{\mathcal{D}}$ § A.6.1			Structural Probes$^{\mathcal{D}}$ § A.6.2	Structural Probes$^{\mathcal{D}}$ § A.6.2	Auxiliary Task$^{\mathcal{D}}$
rules	SEAR$^{\mathcal{M}}$ § A.7.1		Compositional Explanations of Neurons† § A.7.2			

less information → more information

lower abstraction ↓ higher abstraction

	post-hoc: black-box	dataset	gradient	embeddings	white-box	intrinsic: model specific
local explanation						
input features	Occlusion -based § 2.5.2		Gradient -based § 2.5.1			Attenton -based § 2.5.3
adversarial examples	SEA$^{\mathcal{M}}$ § A.1.2		HotFlip § A.1.1			
influential examples		Influence Functions$^{\mathcal{H}}$ § A.2.1 TracIn$^{\mathcal{C}}$ § A.2.3		Representer Pointers† § A.2.2		Prototype Networks
counter-factuals	Polyjuice$^{\mathcal{M},\mathcal{D}}$ § 2.6.1	MiCE$^{\mathcal{M}}$ § 2.6.2				
natural language	predict-then-explain$^{\mathcal{M}}$ § 2.7.2					explain-then-predict$^{\mathcal{M}}$ § 2.7.1
class explanation						
concepts					NIE$^{\mathcal{D}}$ § A.3.1	
global explanation						
vocabulary				Project § A.4.1, Rotate § A.4.2		
ensemble	SP-LIME § A.5.1					
linguistic information	Behavioral Probes$^{\mathcal{D}}$ § A.6.1			Structural Probes$^{\mathcal{D}}$ § A.6.2	Structural Probes$^{\mathcal{D}}$ § A.6.2	Auxiliary Task$^{\mathcal{D}}$
rules	SEAR$^{\mathcal{M}}$ § A.7.1		Compositional Explanations of Neurons† § A.7.2			

Self-explanations

huggingface.co/chat/conversation/65d66

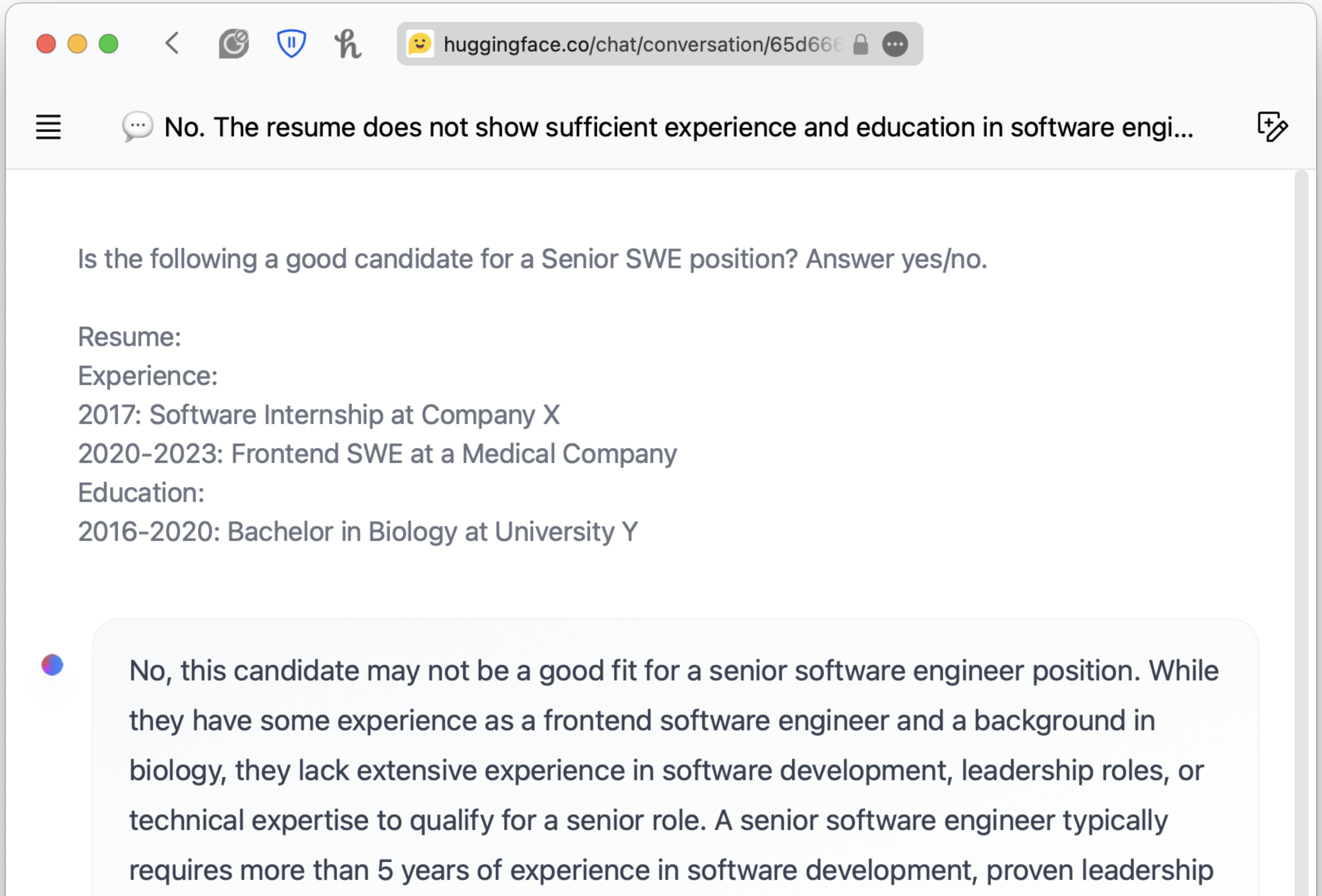
No. The resume does not show sufficient experience and education in software engi...

Is the following a good candidate for a Senior SWE position? Answer yes/no.

Resume:
Experience:
2017: Software Internship at Company X
2020-2023: Frontend SWE at a Medical Company
Education:
2016-2020: Bachelor in Biology at University Y

No, this candidate may not be a good fit for a senior software engineer position. While they have some experience as a frontend software engineer and a background in biology, they lack extensive experience in software development, leadership roles, or technical expertise to qualify for a senior role. A senior software engineer typically requires more than 5 years of experience in software development, proven leadership skills, and a deep understanding of software architecture, algorithms, data structures, and programming languages.

LLMs are high-stakes decision makers



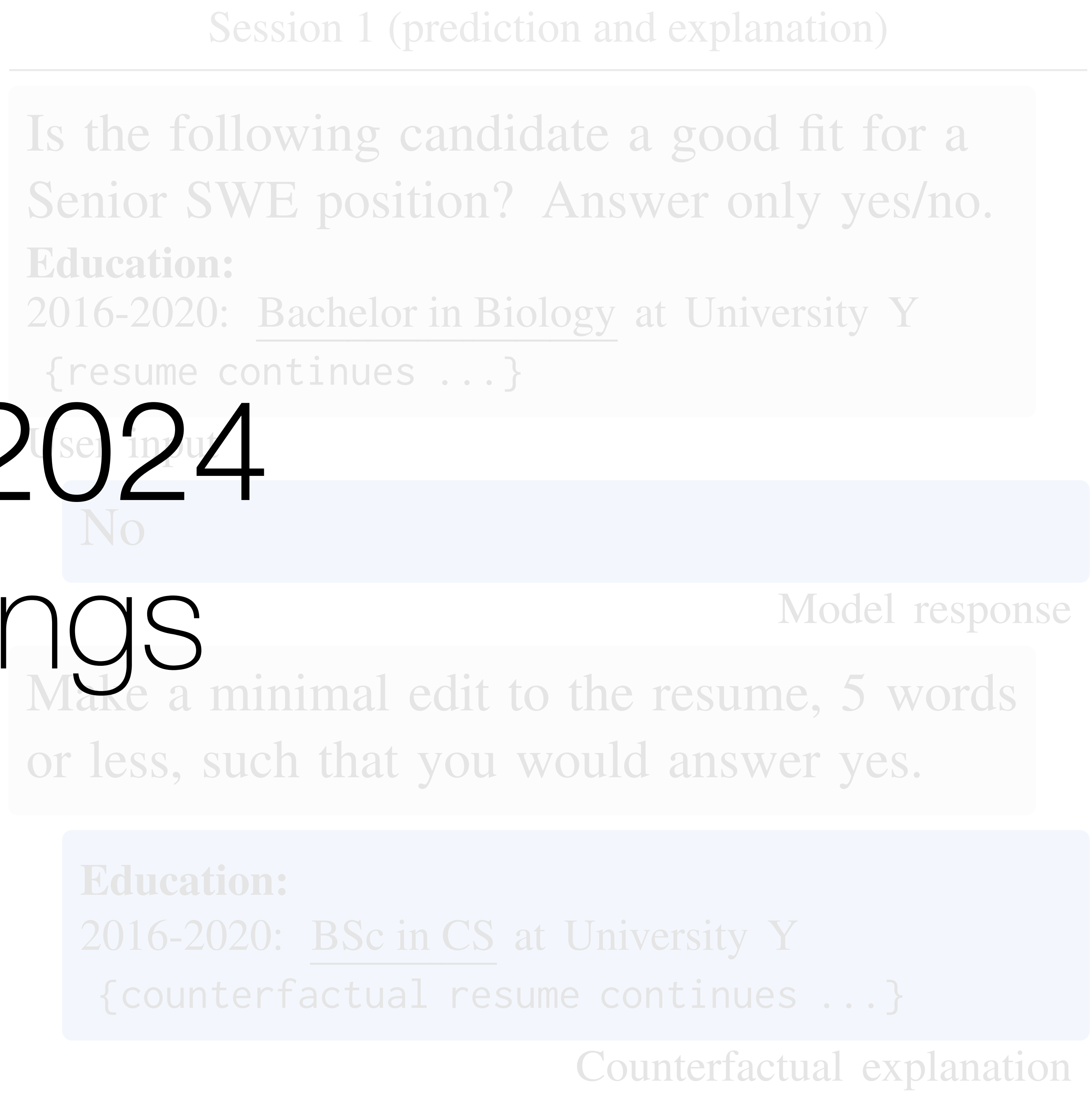
Are self-explanations from Large Language Models faithful?

Andreas Madsen[1,2] **Sarath Chandar**[1,2,4] **Siva Reddy**[1,3,5]
[1] Mila – Quebec AI Institute [2] Polytechnique Montréal [3] McGill University
[4] Canada CIFAR AI Chair [5] Facebook CIFAR AI Chair
{firstname.lastname}@mila.quebec

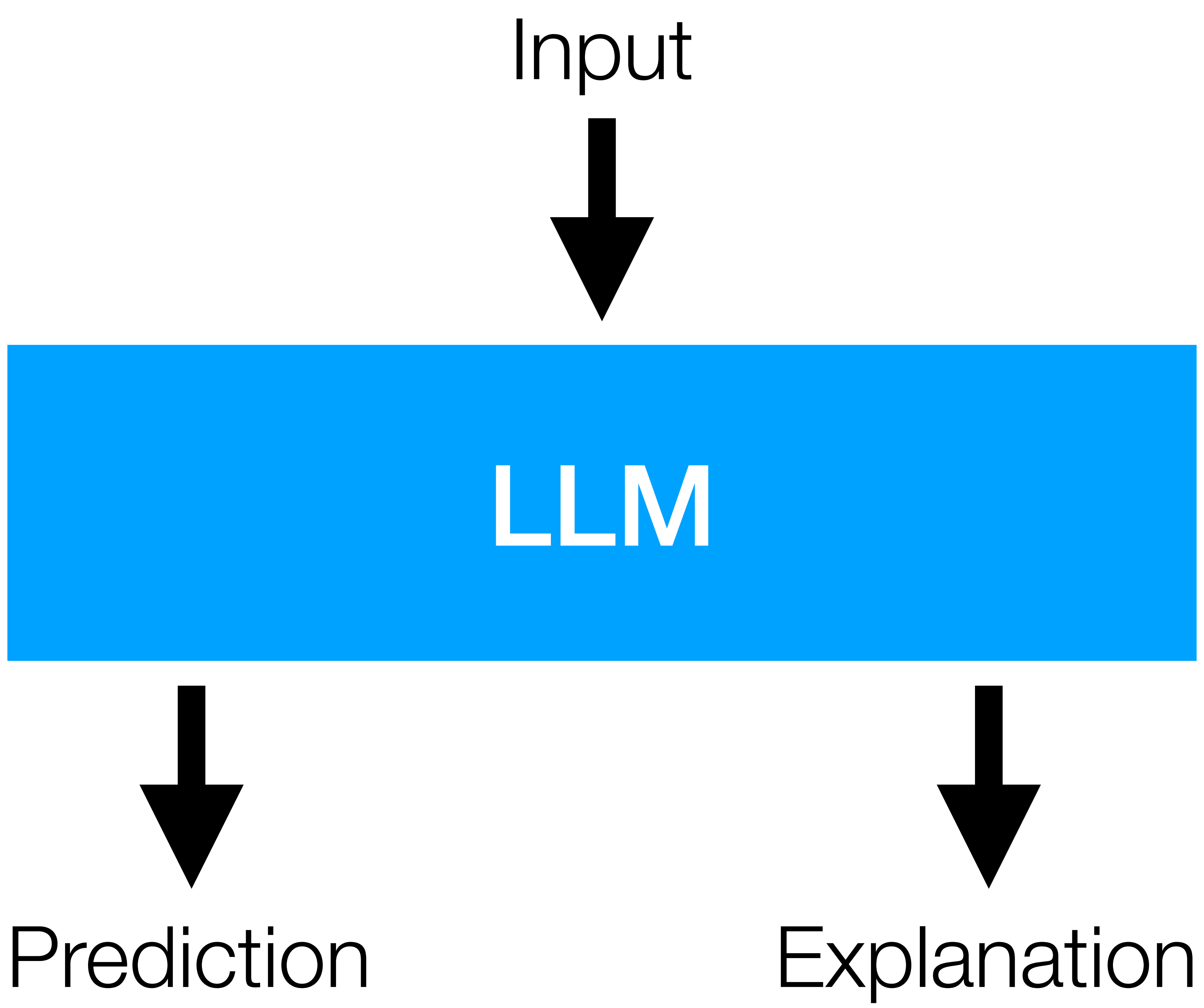
ACL 2024
Findings

Abstract

Instruction-tuned Large Language Models (LLMs) excel at many tasks and will even explain their reasoning, so-called self-explanations. However, convincing and wrong self-explanations can lead to unsupported confidence in LLMs, thus increasing risk. Therefore, it's important to measure if self-explanations truly reflect the model's behavior. Such a measure is called interpretability-faithfulness and is challenging to perform since the ground truth is inaccessible, and many LLMs only have an inference API. To address this, we propose employing self-consistency checks to measure faithfulness. For example, if an LLM says a set of words is important for making a prediction, then it should not be able to make its prediction without these words. While self-



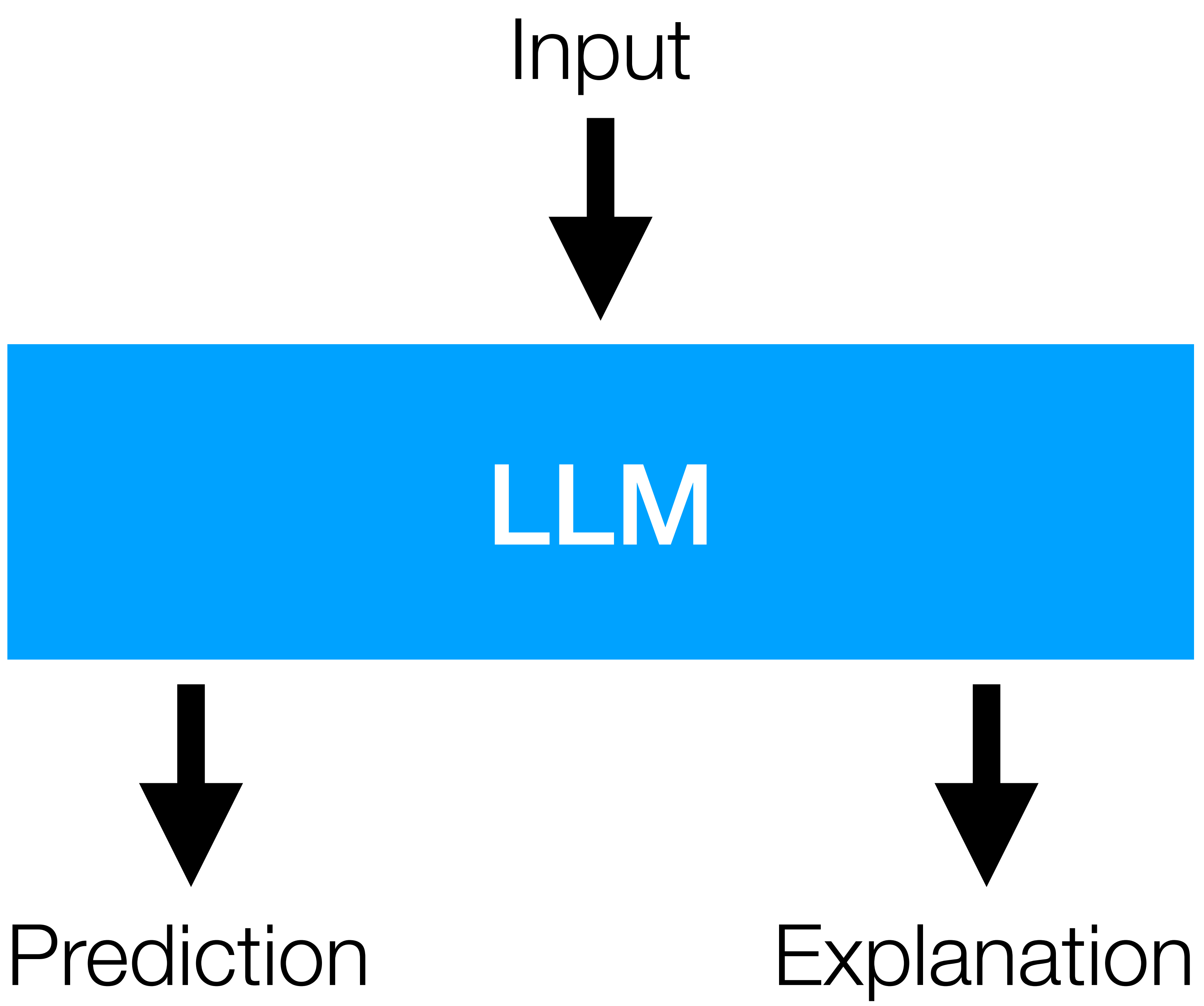
Self-explanations



Self-explanations

Cons

- Explanation is also produced by a black-box.
- Hard to measure faithfulness of free-formed explanations.



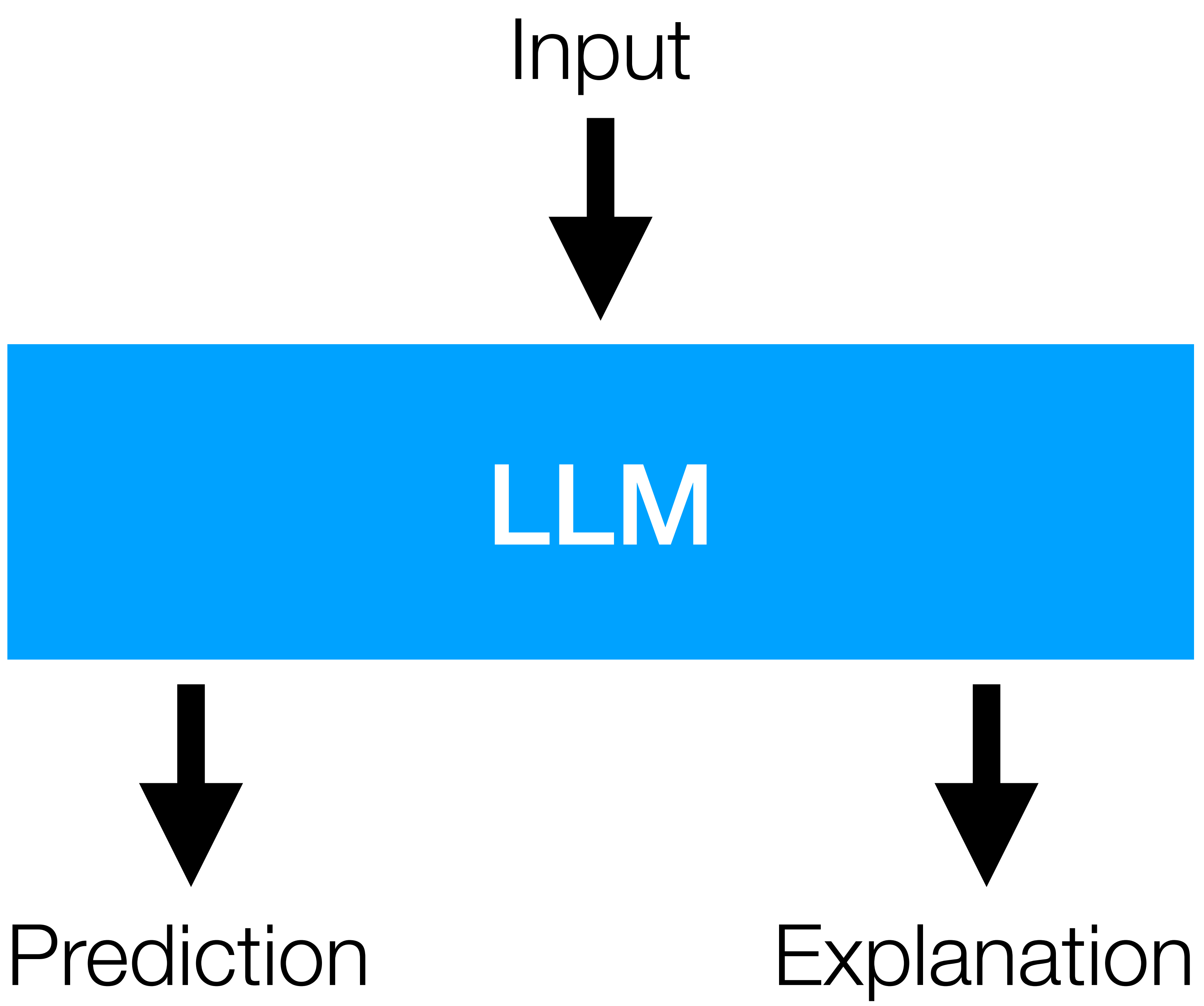
Self-explanations

Cons

- Explanation is also produced by a black-box.
- Hard to measure faithfulness of free-formed explanations.

Pros

- The explanation could have greater access to the prediction-model behavior.



Problem

Hard to measure faithfulness
of free-formed explanations.

Problem

Hard to measure faithfulness
of free-formed explanations.

Solution

Look at a subset of self-explanations
which we know how to validate.

Sentiment counterfactuals

Session 1 (prediction and explanation)

What is the sentiment of the following paragraph? Answer only "positive", "negative", "neutral", or “unknown":

Paragraph: I liked this the movie, it was great.

Positive.

Sentiment counterfactuals

Session 1 (prediction and explanation)

What is the sentiment of the following paragraph? Answer only "positive", "negative", "neutral", or “unknown":

Paragraph: I liked this the movie, it was great.

Positive.

Edit the paragraph such that the sentiment is “negative". Make as few edits as possible.

I disliked this the movie, it was awful.

Sentiment counterfactuals

Session 1 (prediction and explanation)

What is the sentiment of the following paragraph? Answer only "positive", "negative", "neutral", or “unknown":

Paragraph: I liked this the movie, it was great.

Positive.

Edit the paragraph such that the sentiment is “negative". Make as few edits as possible.

I disliked this the movie, it was awful.

Compare

Sentiment counterfactuals

Session 1 (prediction and explanation)

What is the sentiment of the following paragraph? Answer only "positive", "negative", "neutral", or “unknown":

Paragraph: I liked this the movie, it was great.

Positive.

Edit the paragraph such that the sentiment is “negative". Make as few edits as possible.

I disliked this the movie, it was awful.

Copy

Session 2 (Self-consistency)

What is the sentiment of the following paragraph? Answer only "positive", "negative", "neutral", or “unknown":

Paragraph: I disliked this the movie, it was awful.

Negative

Sentiment counterfactuals

Session 1 (prediction and explanation)

What is the sentiment of the following paragraph? Answer only "positive", "negative", "neutral", or “unknown":

Paragraph: I liked this the movie, it was great.

Positive.

Edit the paragraph such that the sentiment is “negative". Make as few edits as possible.

I disliked this the movie, it was awful.

Self-consistent

Faithful

Session 2 (Self-consistency)

What is the sentiment of the following paragraph? Answer only "positive", "negative", "neutral", or “unknown":

Paragraph: I disliked this the movie, it was awful.

Negative

Sentiment feature attribution

Session 1 (prediction and explanation)

What is the sentiment of the following para-
graph? The paragraph can contain redacted
words marked with [REDACTED]. Answer only
"positive", "negative", "neutral", or “unknown":

Paragraph: I liked this the movie, it was great.

No

List the most important words for determining
the sentiment, such that without these words
the sentiment cannot be determined.

Important words: "liked," “great".

Self-consistent

Faithful

Session 2 (Self-consistency)

What is the sentiment of the following para-
graph? The paragraph can contain redacted
words marked with [REDACTED]. Answer only
"positive", "negative", "neutral", or “unknown":

Paragraph: I [REDACTED] this the movie, it
was [REDACTED].

Unknown

Sentiment redaction

Session 1 (prediction and explanation)

What is the sentiment of the following paragraph? The paragraph can contain redacted words marked with [REDACTED]. Answer only "positive", "negative", "neutral", or “unknown":

Paragraph: I liked this the movie, it was great.

No

Redact the most important words for determining the sentiment, by replacing important words with [REDACTED], such that without these words the sentiment can not be determined.

Paragraph: I [REDACTED] this the movie, it was [REDACTED].

Direct redaction

Self-consistent

Faithful

Session 2 (Self-consistency)

What is the sentiment of the following paragraph? The paragraph can contain redacted words marked with [REDACTED]. Answer only "positive", "negative", "neutral", or “unknown":

Paragraph: I [REDACTED] this the movie, it was [REDACTED].

Unknown

Self-consistency checks

The movie was great.

Self-consistency checks

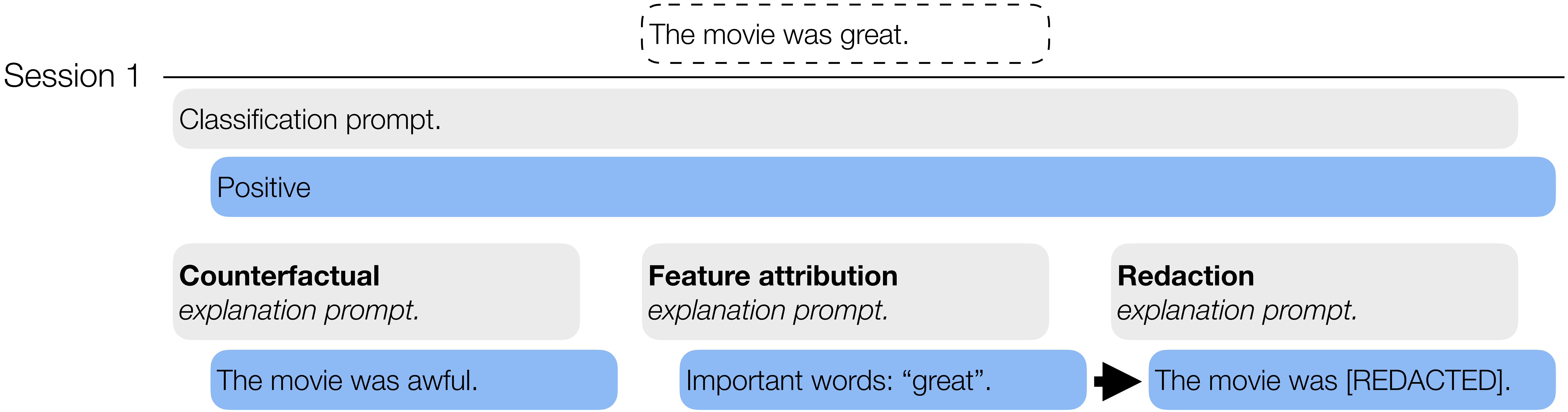
The movie was great.

Session 1

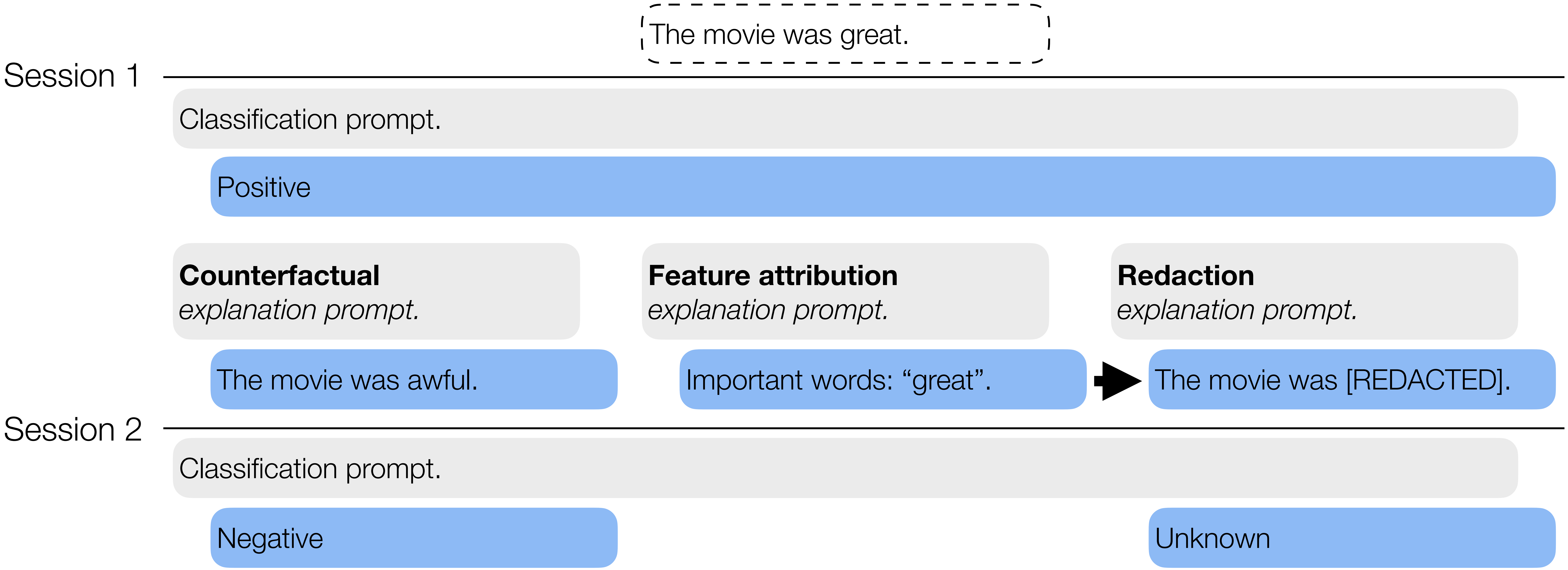
Classification prompt.

Positive

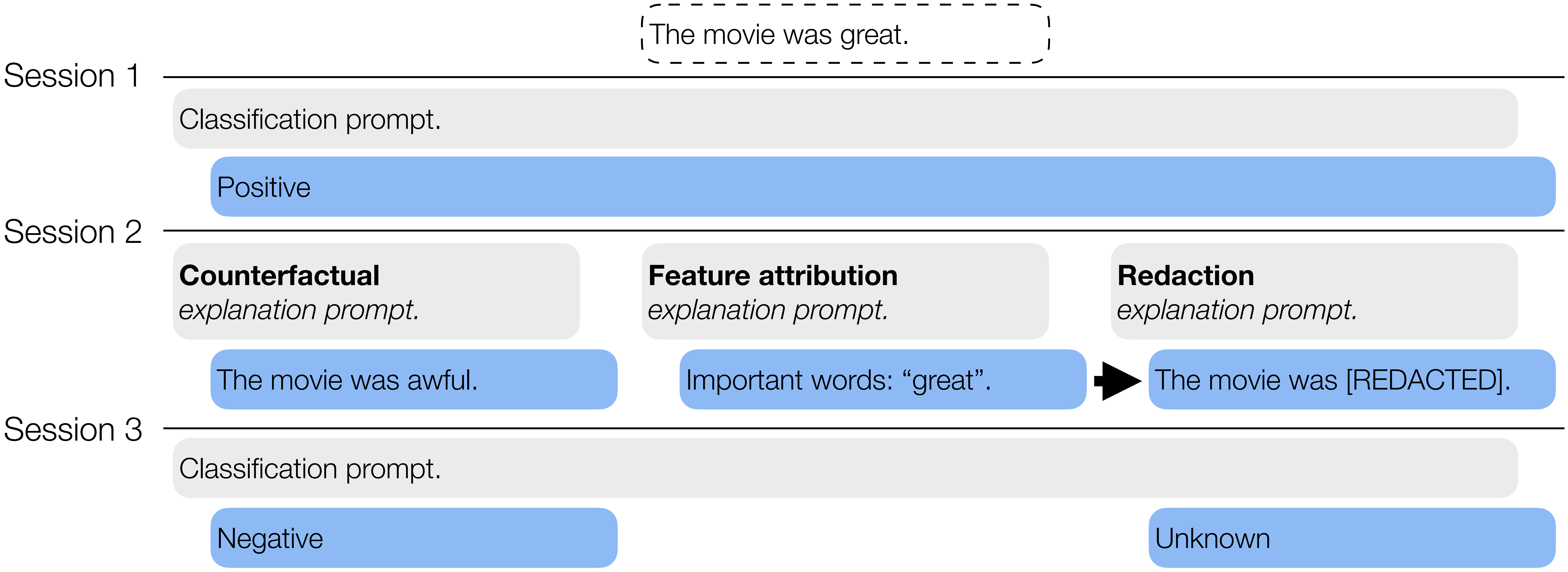
Self-consistency checks



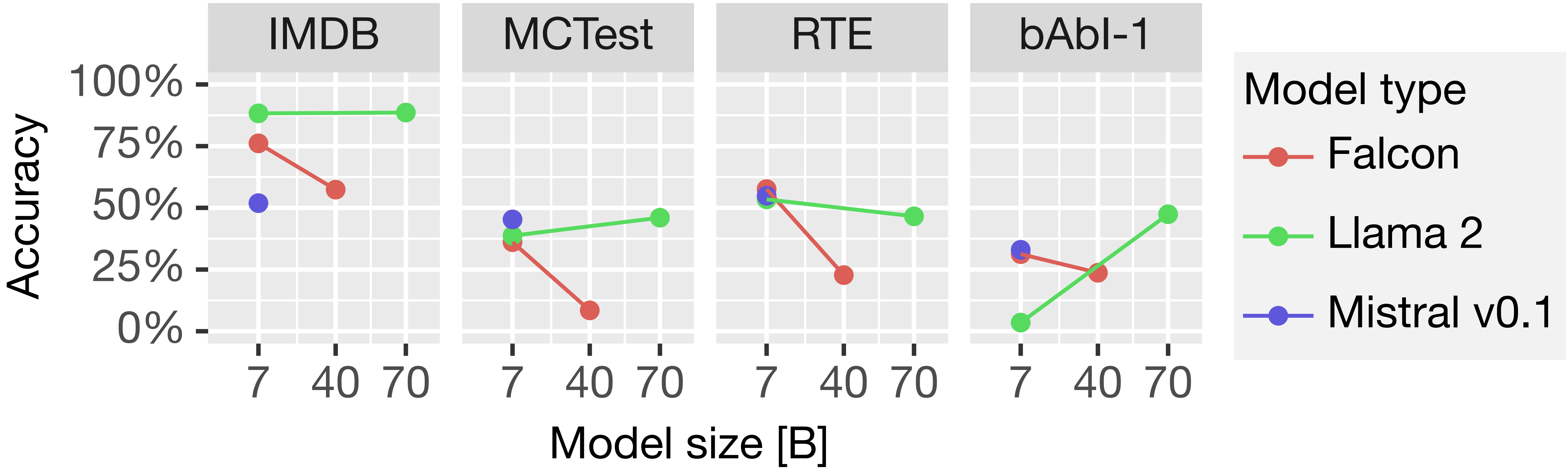
Self-consistency checks



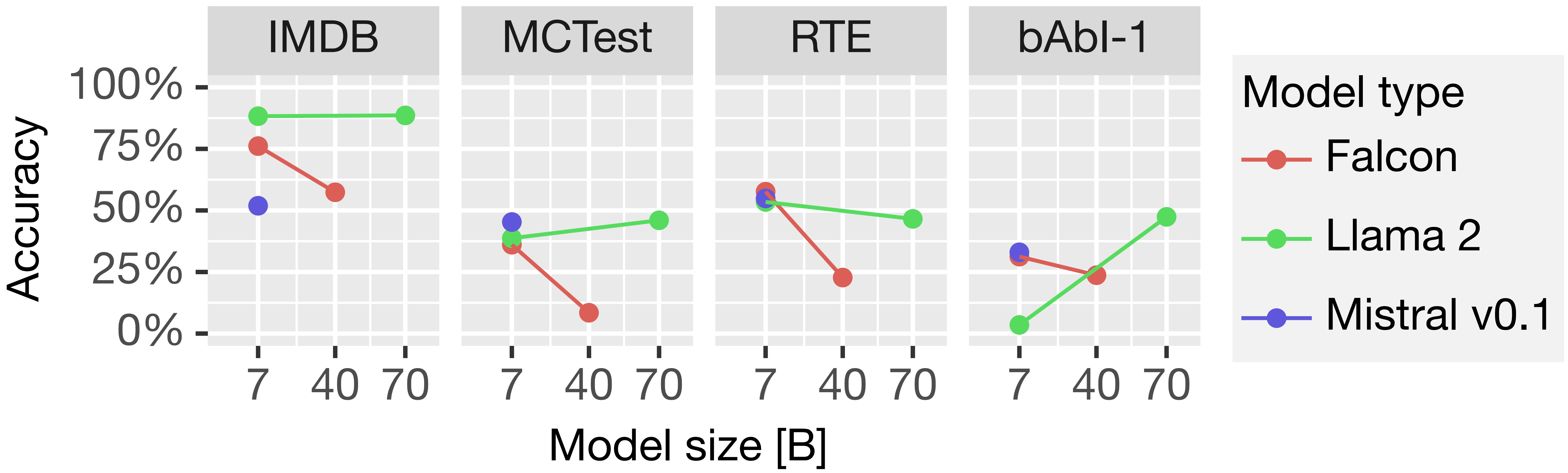
Self-consistency checks



Classification

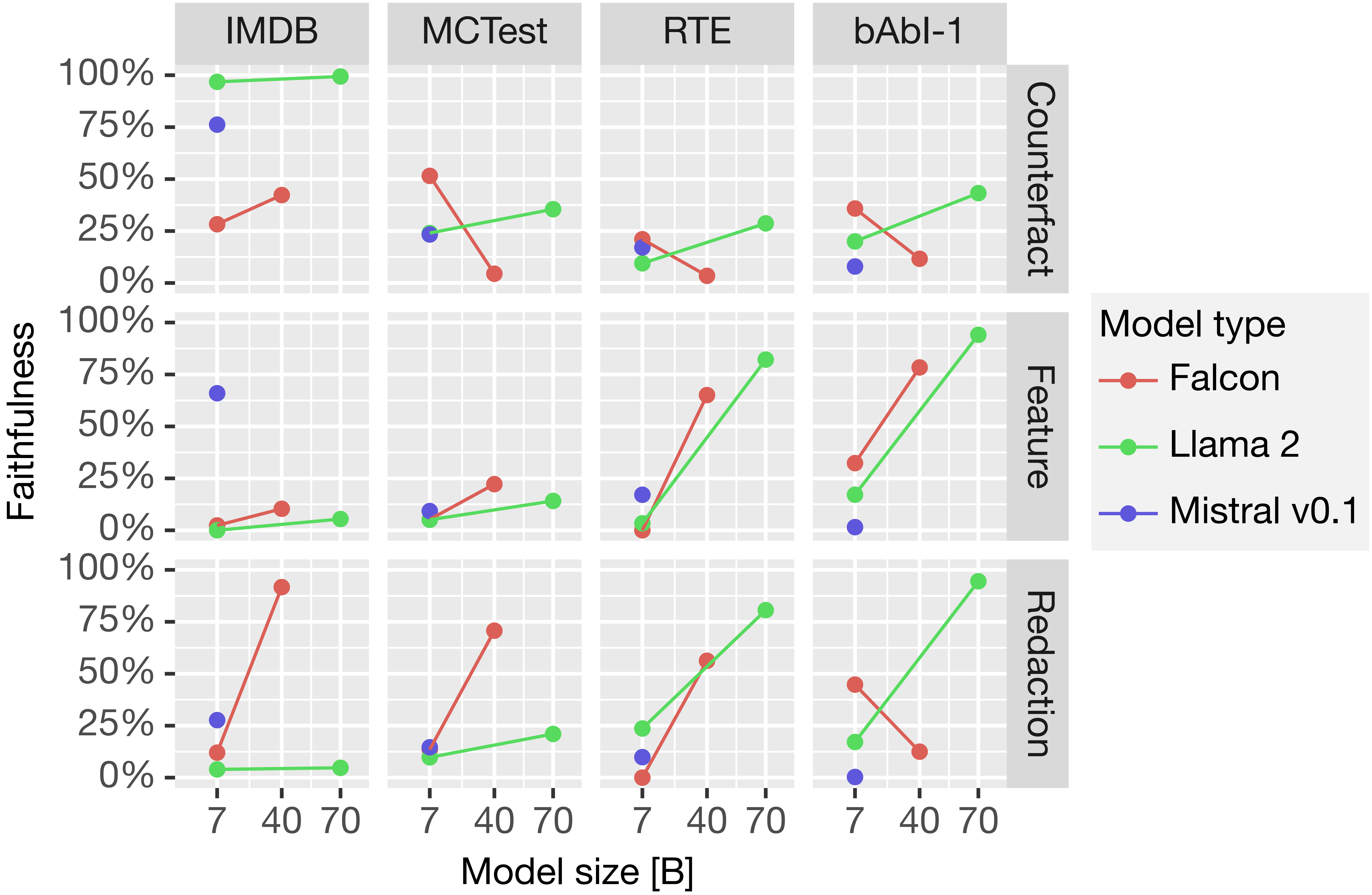


Classification



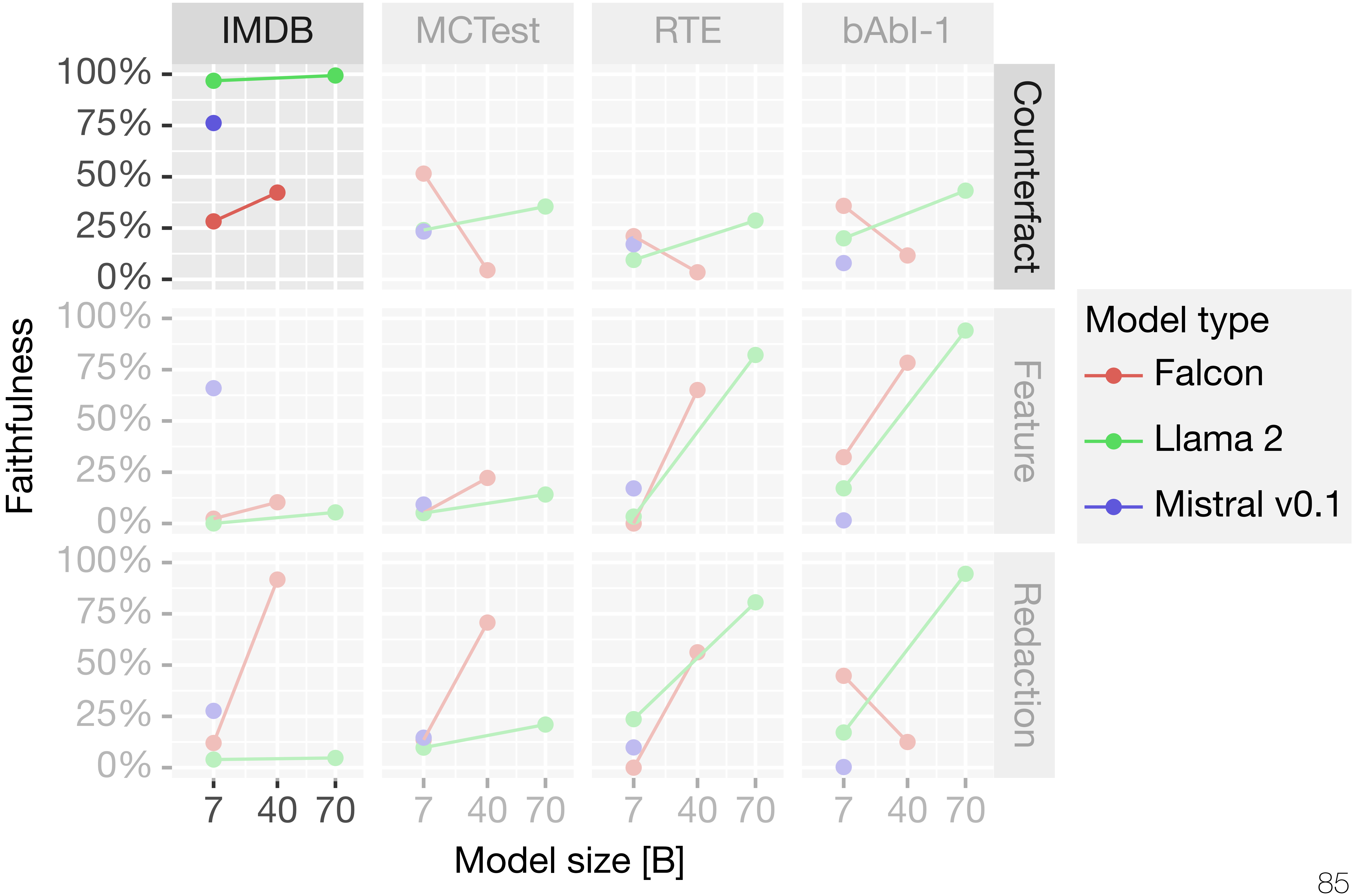
Because the goal is not a high accuracy LLM classifier, we just discard misclassified observations.

Faithfulness



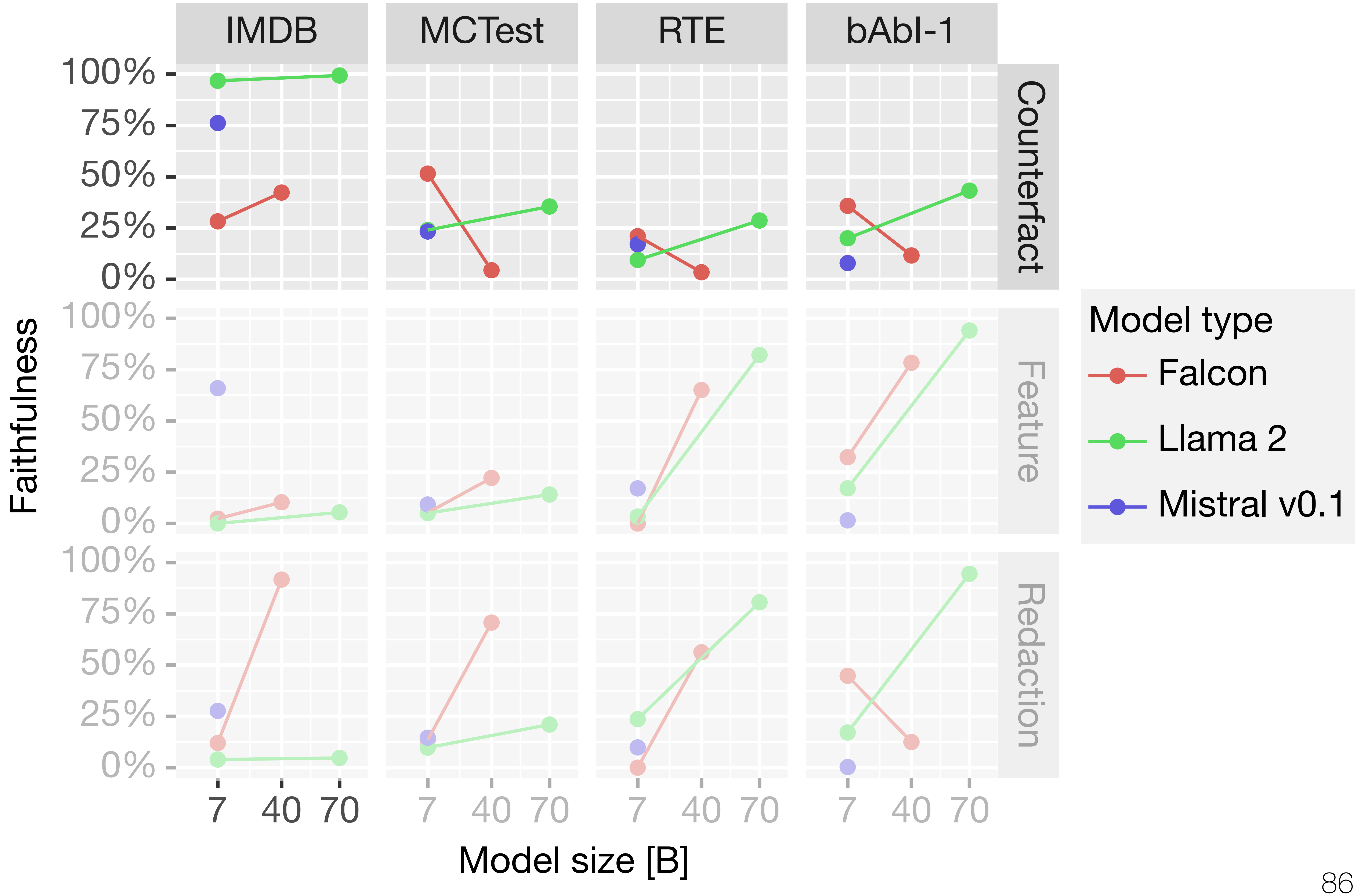
Faithfulness

- Model-dependent.



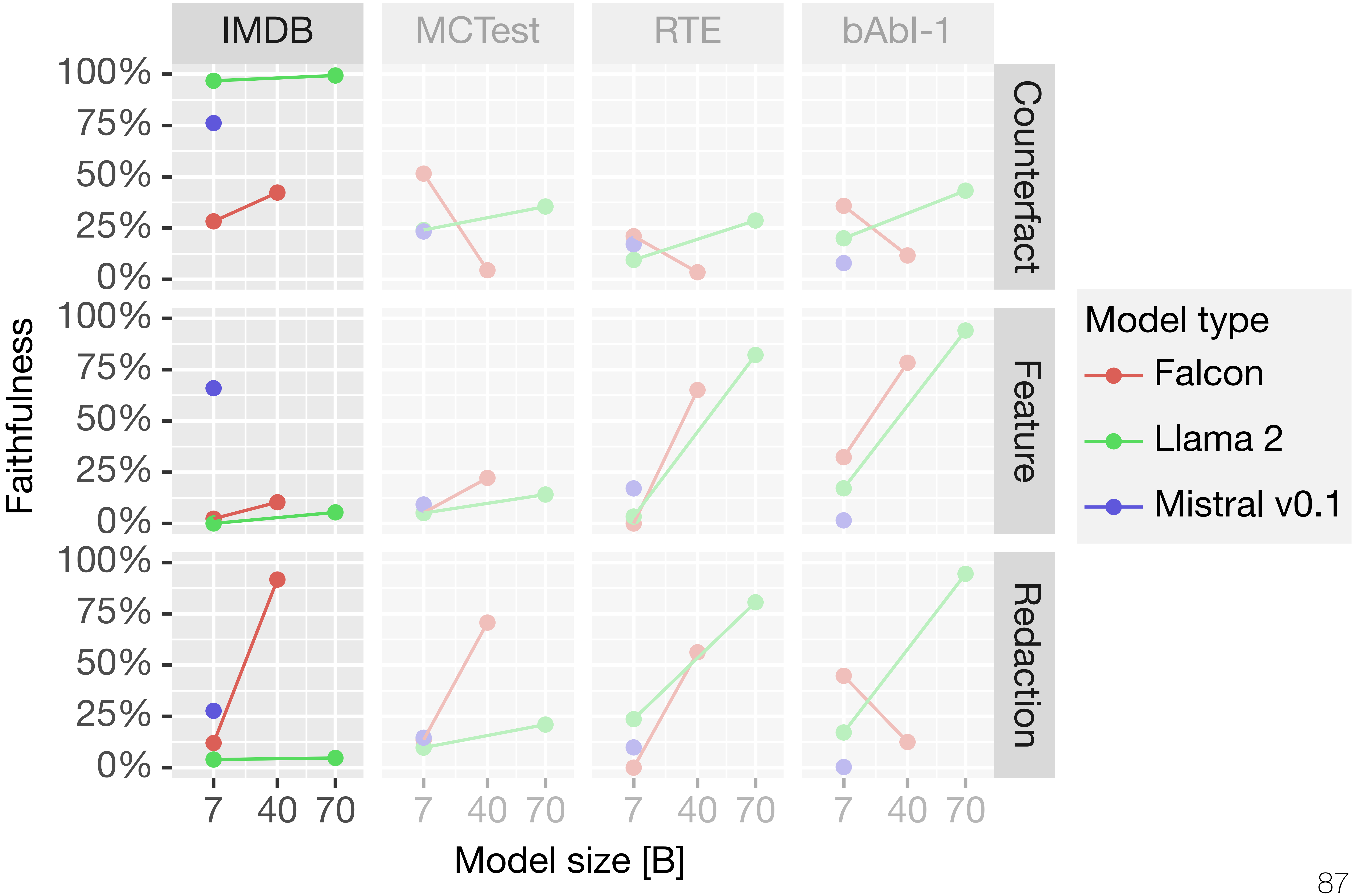
Faithfulness

- Model-dependent.
- Task-dependent.



Faithfulness

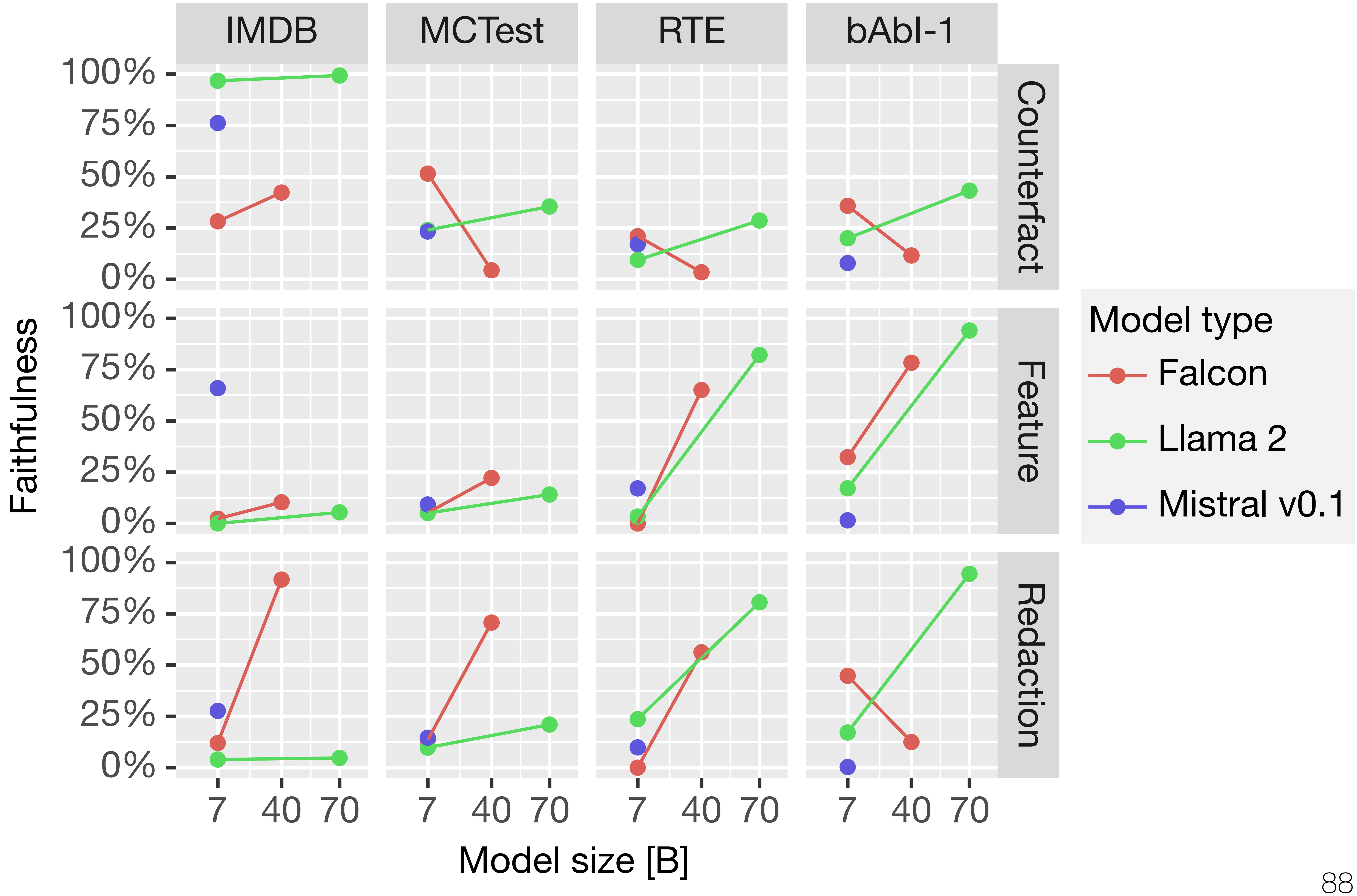
- Model-dependent.
- Task-dependent.
- Explanation-dependent.



Faithfulness

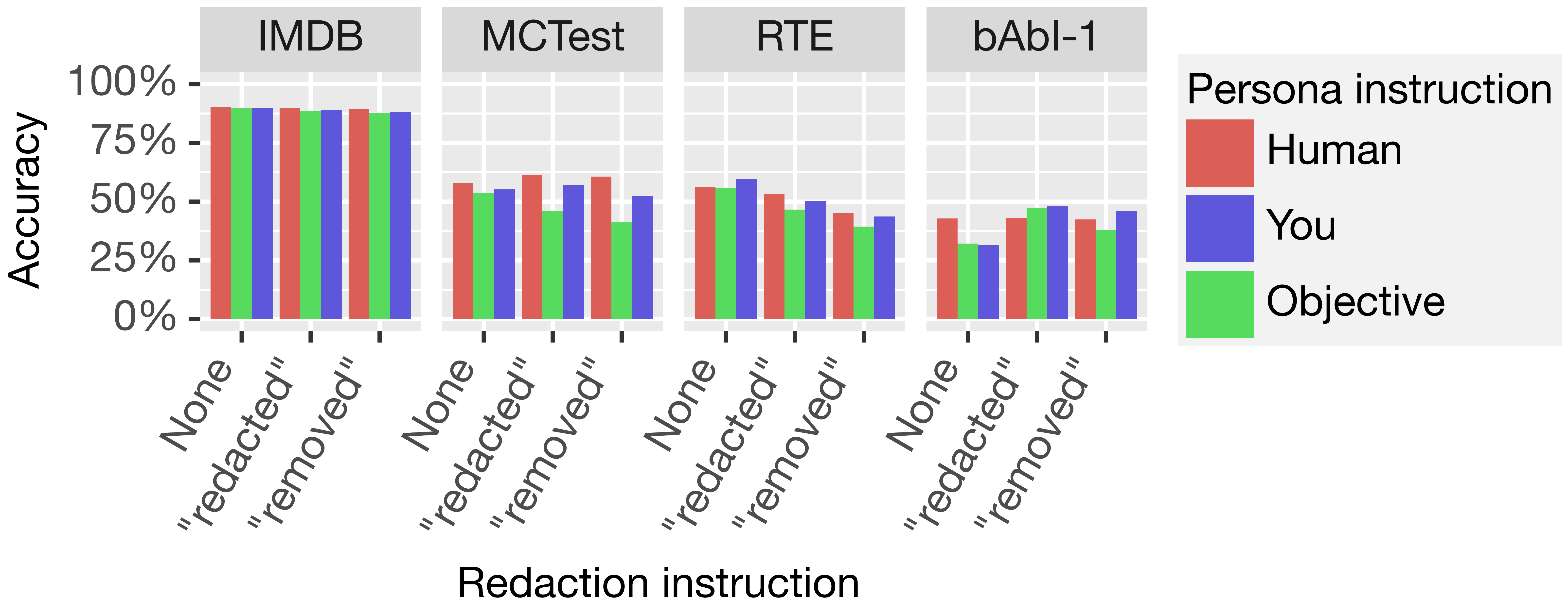
- Model-dependent.
- Task-dependent.
- Explanation-dependent.

In general, we can't trust LLMs' self-explanations.



Robustness

What about prompt variations?



Robustness

If the model was generally faithful
but one prompt variation was not,
that would be problematic.

Robustness

If the model was generally faithful
but one prompt variation was not,
that would be problematic.

How can we make LLMs' self-explanations faithful?

Future work

What are we aligning towards

Human preference.

What are we aligning towards

Human preference.

Humans don't know how the model behaves.

What are we aligning towards

Humans don't know how the model behaves.

Fairwashing

Case 1

Is the following candidate a good fit for a Senior SWE position? Answer only yes/no.

Education:
2016-2020: Bachelor in Biology at University Y

Extra:
Member of Women's Chess Club

No, the education does not match the position.

Case 2

Is the following candidate a good fit for a Senior SWE position? Answer only yes/no.

Education:
2016-2020: Bachelor in Biology at University Y

Extra:
Member of Chess Club

Yes.

[1] Aïvodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., & Tapp, A. Fairwashing: The risk of rationalization. ICML 2019

[2] Aïvodji, U., Arai, H., Gambs, S., & Hara, S. Characterizing the risk of fairwashing, NeurIPS 2021.

Fairwashing

Case 1

Is the following candidate a good fit for a
Senior SWE position? Answer only yes/no.

Education:
2016-2020: Bachelor in Biology at University Y

Extra:
Member of Women's Chess Club

No, the education does not match
the position.

"Preferred"

Case 2

Is the following candidate a good fit for a
Senior SWE position? Answer only yes/no.

Education:
2016-2020: Bachelor in Biology at University Y

Extra:
Member of Women's Chess Club

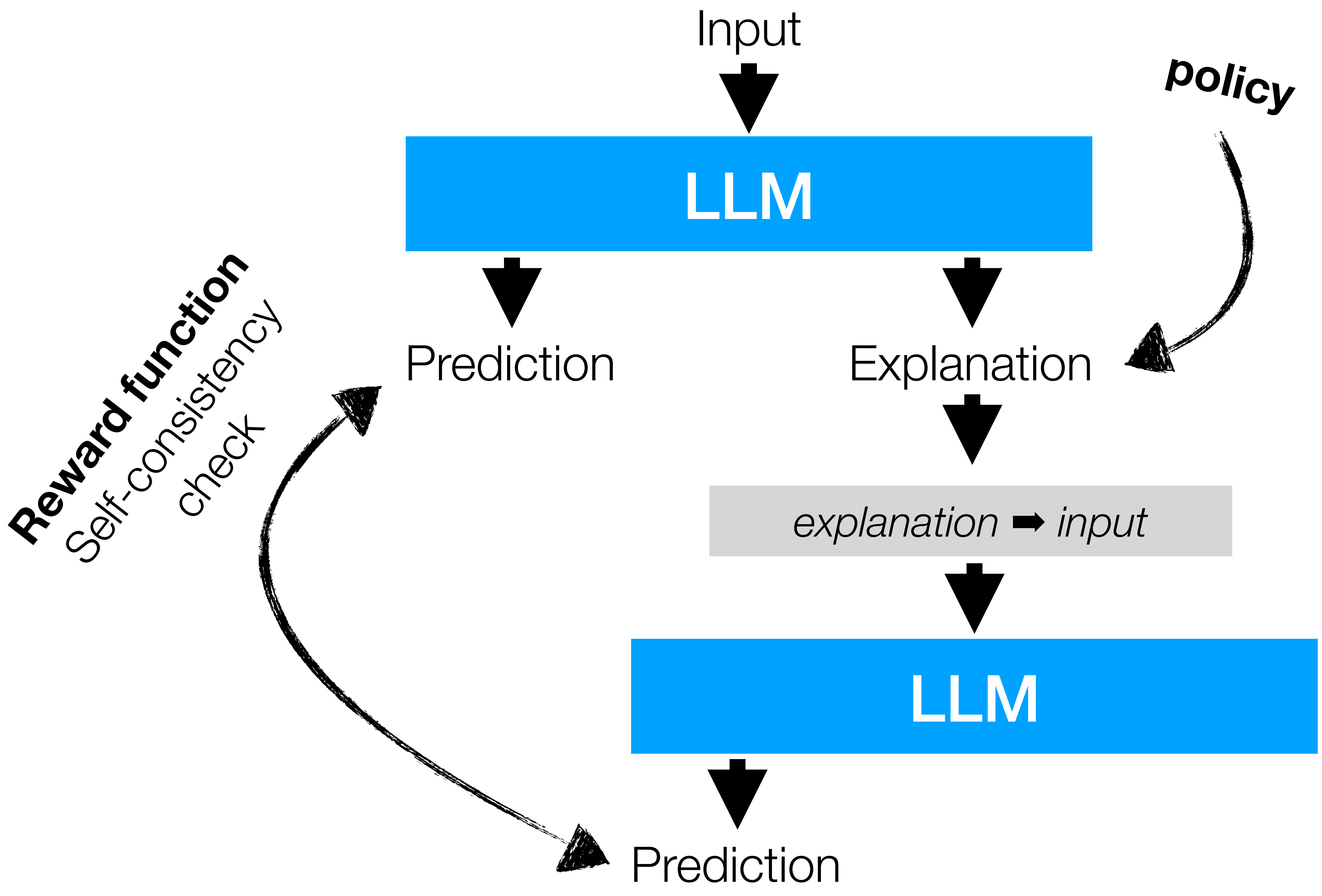
No, because it's a women.

Faithful

[1] Aïvodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., & Tapp, A. Fairwashing: The risk of rationalization. ICML 2019

[2] Aïvodji, U., Arai, H., Gambs, S., & Hara, S. Characterizing the risk of fairwashing, NeurIPS 2021.

Optimizing for faithfulness



Self-explanations

Only models designed to be explained can be explained.

Black-box models are more general purpose.

Future Work

Self-explanations

- Optimize also for faithfulness
- Self-modeling capabilities
- More faithfulness metrics

Faithfulness Measurable Models

- Applies to CLMs
- Better Optimizations
- Applies to other explanations

Conclusion

How to provide and ensure faithful explanations for complex general-purpose neural NLP models?

Research question

This question can be answered:

- By developing **new paradigms** that design models to be explained without employing architectural constraints.
- By focusing on developing **accurate faithfulness metrics**.
- By focusing on **importance measures** that have had a notoriously troubling history regarding faithfulness.
- By taking advantage of properties specific to natural language and **NLP** models.

Research hypothesis

How to provide and ensure faithful explanations for complex general-purpose neural NLP models?

Research question

This question can be answered:

- By developing **new paradigms** that design models to be explained without employing architectural constraints.
- By focusing on developing **accurate faithfulness metrics**.
- By focusing on **importance measures** that have had a notoriously troubling history regarding faithfulness.
- By taking advantage of properties specific to natural language and **NLP** models.

Research hypothesis

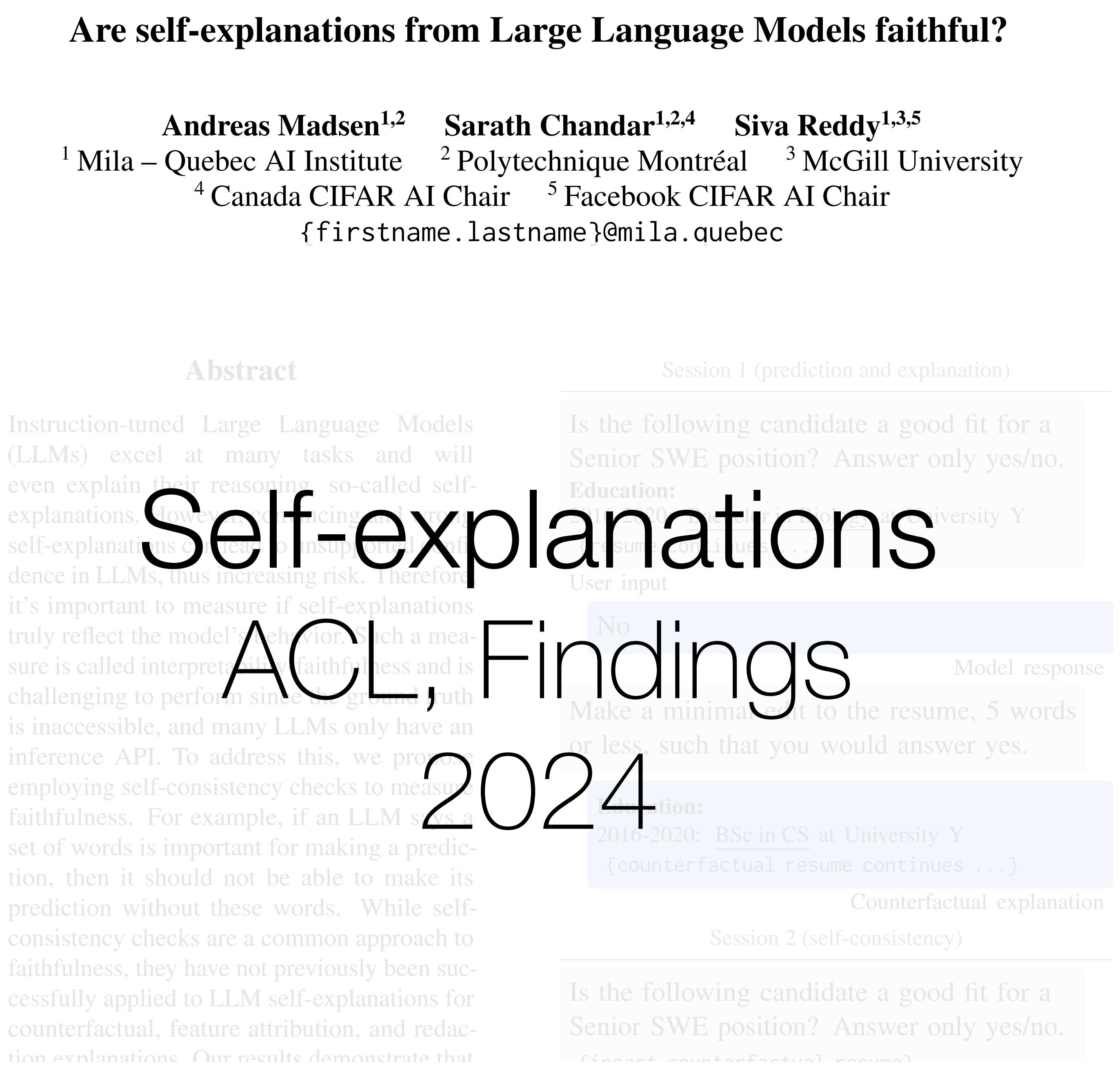
Model and task-dependent faithfulness

- The faithfulness of post-hoc and attention is model and task-dependent.
- Shown on importance measures and self-explanations. Simultaneously works [1,2] with same conclusion.

Evaluating the Faithfulness of Importance Measures in NLP by Recursively Masking Allegedly Important Tokens and Retraining

Andreas Madsen[1,2] Nicholas Meade[1,3,*] Vaibhav Adlakha[1,3,*] Siva Reddy[1,3,4]
[1] Mila – Quebec AI Institute [2] Polytechnique Montréal
[3] McGill University [4] Facebook CIFAR AI Chair
{firstname.lastname}@mila.quebec

Recursive ROAR
EMNLP, Findings
2022



Are self-explanations from Large Language Models faithful?

Andreas Madsen[1,2] Sarath Chandar[1,2,4] Siva Reddy[1,3,5]
[1] Mila – Quebec AI Institute [2] Polytechnique Montréal [3] McGill University
[4] Canada CIFAR AI Chair [5] Facebook CIFAR AI Chair
{firstname.lastname}@mila.quebec

Self-explanations
ACL, Findings
2024

Same conclusion in: [1] Bastings, J., et al. "Will You Find These Shortcuts?" A Protocol for Evaluating the Faithfulness of Input Salience Methods for Text Classification. EMNLP 2022

[2] Lanham, T., et al. Measuring Faithfulness in Chain-of-Thought Reasoning. Pre-print 2023.

Model and task-dependent faithfulness

- The faithfulness of post-hoc and attention is model and task-dependent.
- Shown on importance measures and self-explanations. Simultaneously works [1,2] with same conclusion.
- Likely to explain why there is so much debate on is X-method faithful.

Evaluating the Faithfulness of Importance Measures in NLP by Recursively Masking Allegedly Important Tokens and Retraining

Andreas Madsen[1,2] Nicholas Meade[1,3,*] Vaibhav Adlakha[1,3,*] Siva Reddy[1,3,4]
[1] Mila – Quebec AI Institute [2] Polytechnique Montréal
[3] McGill University [4] Facebook CIFAR AI Chair
{firstname.lastname}@mila.quebec

Recursive ROAR
EMNLP, Findings 2022

Are self-explanations from Large Language Models faithful?

Andreas Madsen[1,2] Sarath Chandar[1,2,4] Siva Reddy[1,3,5]
[1] Mila – Quebec AI Institute [2] Polytechnique Montréal [3] McGill University
[4] Canada CIFAR AI Chair [5] Facebook CIFAR AI Chair
{firstname.lastname}@mila.quebec

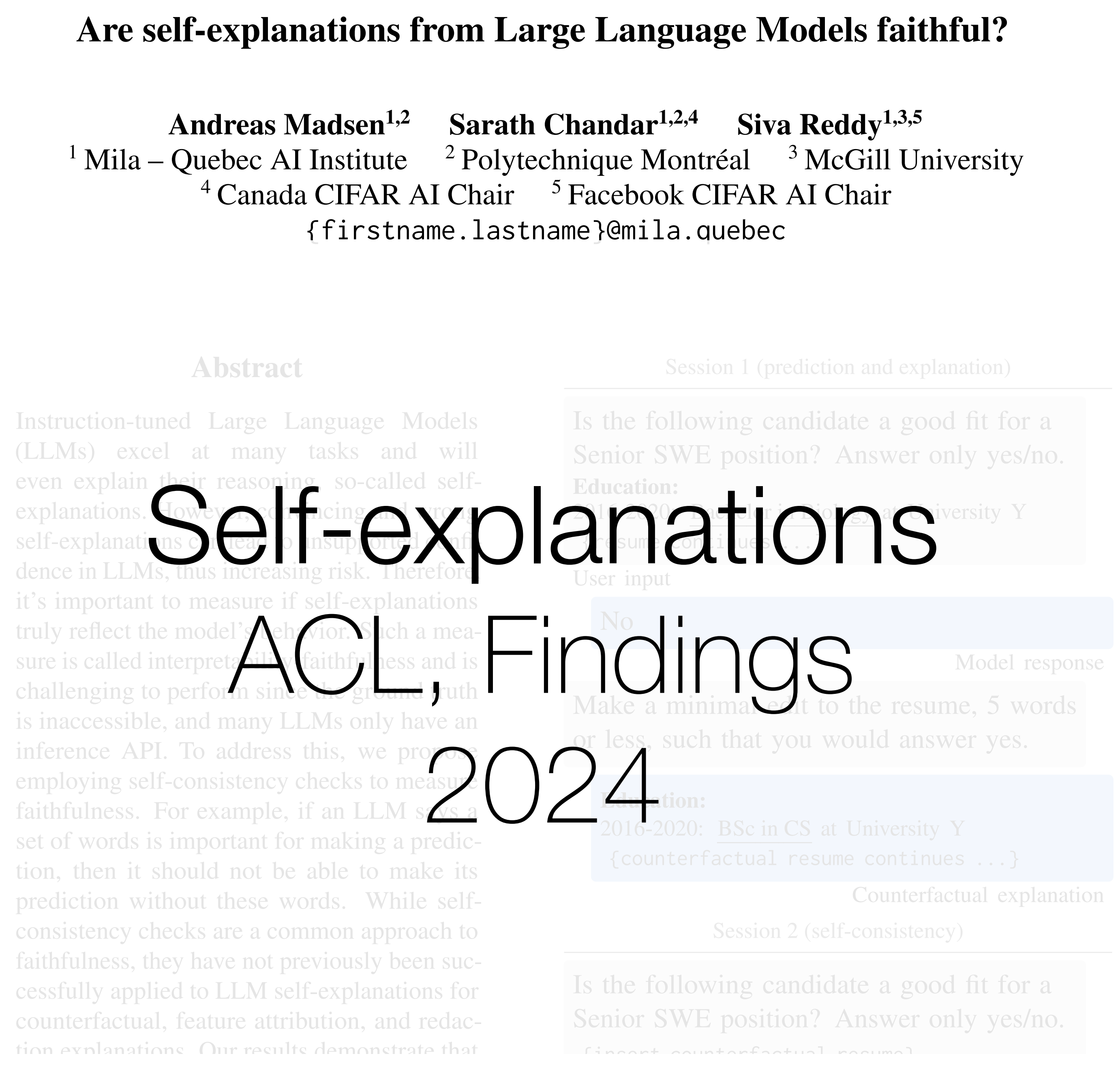
Self-explanations
ACL, Findings 2024

Same conclusion in: [1] Bastings, J., et al. "Will You Find These Shortcuts?" A Protocol for Evaluating the Faithfulness of Input Salience Methods for Text Classification. EMNLP 2022

[2] Lanham, T., et al. Measuring Faithfulness in Chain-of-Thought Reasoning. Pre-print 2023.

Model and task-dependent faithfulness

- The faithfulness of post-hoc and attention is model and task-dependent.
- Shown on importance measures and self-explanations. Simultaneously works [1,2] with same conclusion.
- Likely to explain why there is so much debate on is X-method faithful.
- Only revealed using sufficiently accurate faithfulness metric at large scope.



Same conclusion in: [1] Bastings, J., et al. "Will You Find These Shortcuts?" A Protocol for Evaluating the Faithfulness of Input Salience Methods for Text Classification. EMNLP 2022

[2] Lanham, T., et al. Measuring Faithfulness in Chain-of-Thought Reasoning. Pre-print 2023.

Consistent faithfulness

Recursive ROAR

Model and task-dependent

↓

Faithfulness Measurable Models

Masked fine-tuning creates consistently faithful explanations.

Self-explanations faithfulness metric

Explanation, model and task-dependent

↓

Faithfulness as a reward function

?

How to provide and ensure faithful explanations for complex general-purpose neural NLP models?

Research question

This question can be answered:

- By developing **new paradigms** that design models to be explained without employing architectural constraints.
- By focusing on developing **accurate faithfulness metrics**.
- By focusing on **importance measures** that have had a notoriously troubling history regarding faithfulness.
- By taking advantage of properties specific to natural language and **NLP** models.

Research hypothesis

New Interpretability Paradigms

Faithfulness measurable models

Model is designed such that measuring faithfulness is easy.

Self-explanations

Model is designed such that it can explain itself.

Only models designed to be explained can be explained.

Black-box models are more general purpose.

Conclusion

- The faithfulness of post-hoc methods is **model and task-dependent**.
- Yes, It's possible to develop **new interpretability paradigms**, which show consistent faithfulness.

Pitfalls and Principles

Principles

The two options for measuring faithfulness:

a) Use an intervention, but avoid out-of-distribution issues.
b) Use a ground truth, but make sure it's an actual ground truth.

Pitfalls

a) If correlating, it must be done with a known faithful explanation (which likely doesn't exist).
b) Don't assume the model is reasonable (or accurate?).
c) Don't assume you know what correct explanation looks like (follows previous).
d) Don't mutate the internals of a model to validate explanation, you may escape the manifold.
e) Don't probe the model behavior with out-of-distribution data.
f) Don't use a different model to comment about the original model, unless the model behavior is identical.
g) Don't assume faithfulness generalize to other datasets or models without validation.
h) Not declaring what faithfulness measures. For example, gradient is faithful it is just not a measure of importance.
i) Thinking there is just one correct explanation (importance measure) without a mathematical proof of uniqueness.

Explanation-interpretation gap

Explanation-interpretation gap

Mathematical/axiomatic faithfulness metric.

e.g. “How good is the gradient approximation?”, “are they Sharply values?”

Faithfulness metric that reflects how we communicate explanations to humans.

e.g. “Without this token the model prediction changes significantly”

All the gaps

Interference error + leakage error

Approximation error

Explanation-interpretation gap

Actual issues	Approximations	Mathematical/axiomatic faithfulness metric.	Faithfulness metric that reflects how we communicate explanations to humans.
e.g. $\mathbf{x} \odot \nabla_x f(\mathbf{x})$	*e.g. Sharply approximation.*	*e.g. "How good is the gradient approximation?", "are they Sharply values?"*	*e.g. "Without this token the model prediction changes significantly"*

Survey

less information → more information

lower abstraction → higher abstraction

<table>
<tr><th></th><th colspan="5">post-hoc</th><th>intrinsic</th></tr>
<tr><th></th><th>black-box</th><th>dataset</th><th>gradient</th><th>embeddings</th><th>white-box</th><th>model specific</th></tr>
<tr><td colspan="7">local explanation</td></tr>
<tr><td>input features</td><td>Occlusion -based § 2.5.2</td><td></td><td>Gradient -based § 2.5.1</td><td></td><td></td><td>Attenton -based § 2.5.3</td></tr>
<tr><td>adversarial examples</td><td>SEA$^{\mathcal{M}}$ § A.1.2</td><td></td><td colspan="2">HotFlip § A.1.1</td><td></td><td></td></tr>
<tr><td>influential examples</td><td></td><td colspan="2">Influence Functions$^{\mathcal{H}}$ § A.2.1
TracIn$^{\mathcal{C}}$ § A.2.3</td><td colspan="2">Representer Pointers† § A.2.2</td><td>Prototype Networks</td></tr>
<tr><td>counter-factuals</td><td>Polyjuice$^{\mathcal{M},\mathcal{D}}$ § 2.6.1</td><td colspan="2">MiCE$^{\mathcal{M}}$ § 2.6.2</td><td></td><td></td><td></td></tr>
<tr><td>natural language</td><td>predict-then-explain$^{\mathcal{M}}$ § 2.7.2</td><td></td><td></td><td></td><td></td><td>explain-then-predict$^{\mathcal{M}}$ § 2.7.1</td></tr>
<tr><td colspan="7">class explanation</td></tr>
<tr><td>concepts</td><td></td><td></td><td></td><td></td><td>NIE$^{\mathcal{D}}$ § A.3.1</td><td></td></tr>
<tr><td colspan="7">global explanation</td></tr>
<tr><td>vocabulary</td><td></td><td></td><td></td><td>Project § A.4.1,
Rotate § A.4.2</td><td></td><td></td></tr>
<tr><td>ensemble</td><td colspan="2">SP-LIME § A.5.1</td><td></td><td></td><td></td><td></td></tr>
<tr><td>linguistic information</td><td>Behavioral Probes$^{\mathcal{D}}$ § A.6.1</td><td></td><td></td><td>Structural Probes$^{\mathcal{D}}$ § A.6.2</td><td>Structural Probes$^{\mathcal{D}}$ § A.6.2</td><td>Auxiliary Task$^{\mathcal{D}}$</td></tr>
<tr><td>rules</td><td>SEAR$^{\mathcal{M}}$ § A.7.1</td><td colspan="4">Compositional Explanations of Neurons† § A.7.2</td><td></td></tr>
</table>

Input features

Local Explanation

		$p(y\|\mathbf{x})$	y	c
$\mathbf{x}$	the year 's best and most unpredictable comedy	0.91	1	1
$\mathbf{x}$	we never feel anything for these characters	0.95	0	0
$\mathbf{x}$	handsome but unfulfilling suspense drama	0.18	0	1

Which tokens are most important for the prediction?

Adversarial examples

Local Explanation

		$p(y\|\mathbf{x})$	y
$\mathbf{x}$	the year 's best and most unpredictable comedy	0.91	1
	the year 's finest and most unpredictable comedy	0.30	-
$\tilde{\mathbf{x}}$	the year 's finest and most unforeseeable comedy	0.08	-
$\mathbf{x}$	we never feel anything for these characters	0.95	0
$\tilde{\mathbf{x}}$	we never feel anything for these people	0.03	-

What would break the model's prediction?

Influential examples

Local Explanation

	$\mathbf{x}$	$p(y\|\mathbf{x})$	y	Δ
$\mathbf{x}$	`the year 's best and most unpredictable comedy`	0.91	1	-
$\tilde{\mathbf{x}}$	`a delightfully unpredictable , hilarious comedy`	0.95	1	3.82
$\tilde{\mathbf{x}}$	`loud and thoroughly obnoxious comedy`	0.98	0	-1.51

What training examples influenced the prediction?

Counterfactuals

Local Explanation

	$\mathbf{x}$	$p(y\|\mathbf{x})$	y
$\mathbf{x}$	the year 's best and most unpredictable comedy	0.91	1
	the year 's worst and most unpredictable comedy	0.59	-
$\tilde{\mathbf{x}}$	the year 's worst and most predictable comedy	0.04	-
$\mathbf{x}$	we never feel anything for these characters	0.95	0
	we can feel anything for these characters	0.73	-
$\tilde{\mathbf{x}}$	we can feel anything for these animals	0.01	-

What does the model consider a valid opposite example?

Natural Language

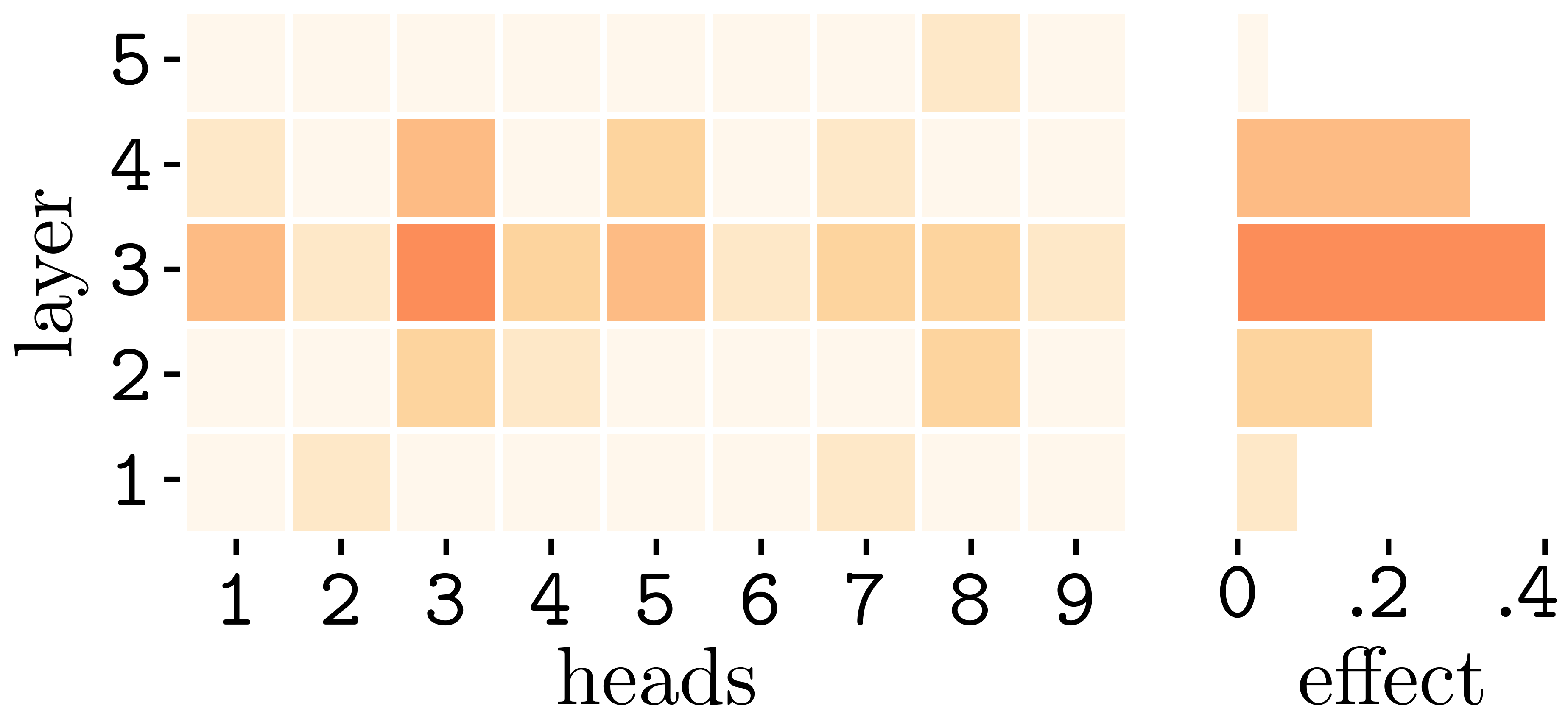
Local Explanation

	$\mathbf{x}$	$p(y\|\mathbf{x})$	y
$\mathbf{x}$	`the` `year` `'s` `best` `and` `most` `unpredictable` `comedy`	0.91	1
	unpredictable comedies are funny	-	-
$\mathbf{x}$	`we` `never` `feel` `anything` `for` `these` `characters`	0.95	0
	it is important to feel for characters	-	-

What would a generated natural language explanation be?

Concepts

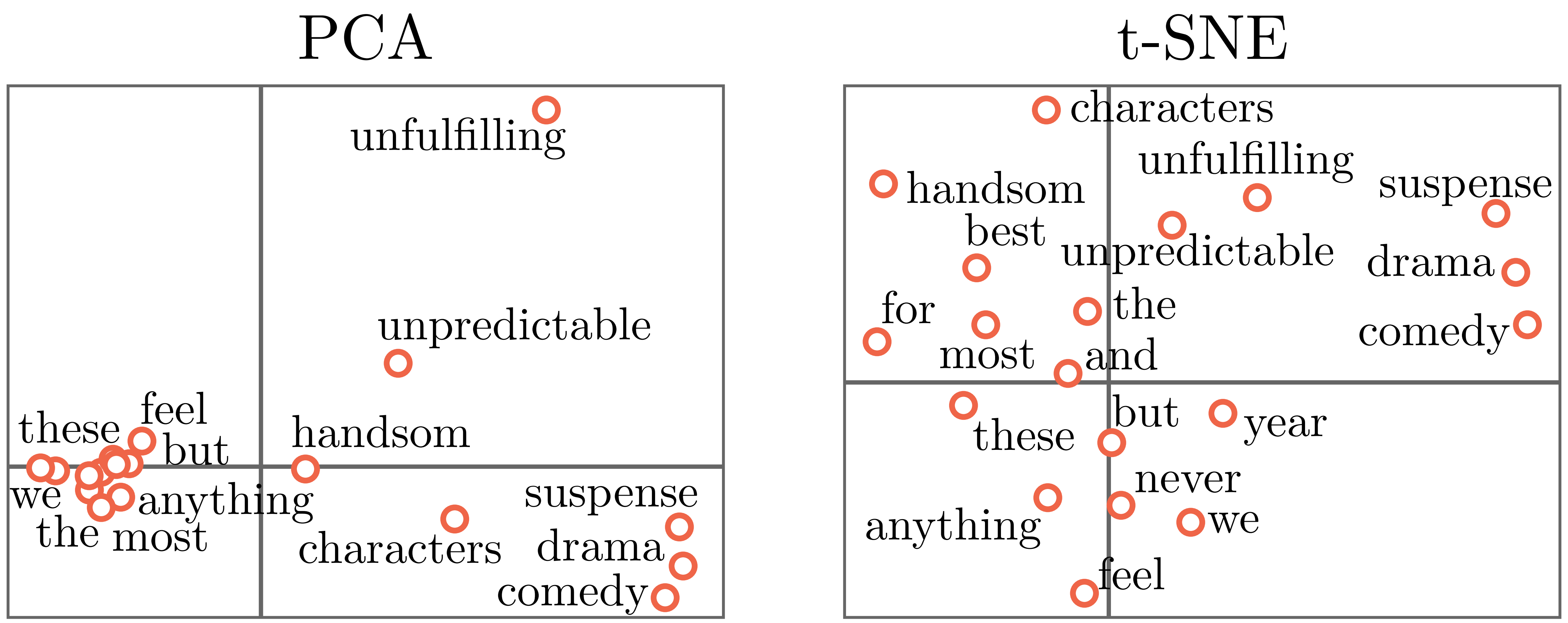
Class Explanation



What concepts (e.g. occupations) can explain a class?

Vocabulary

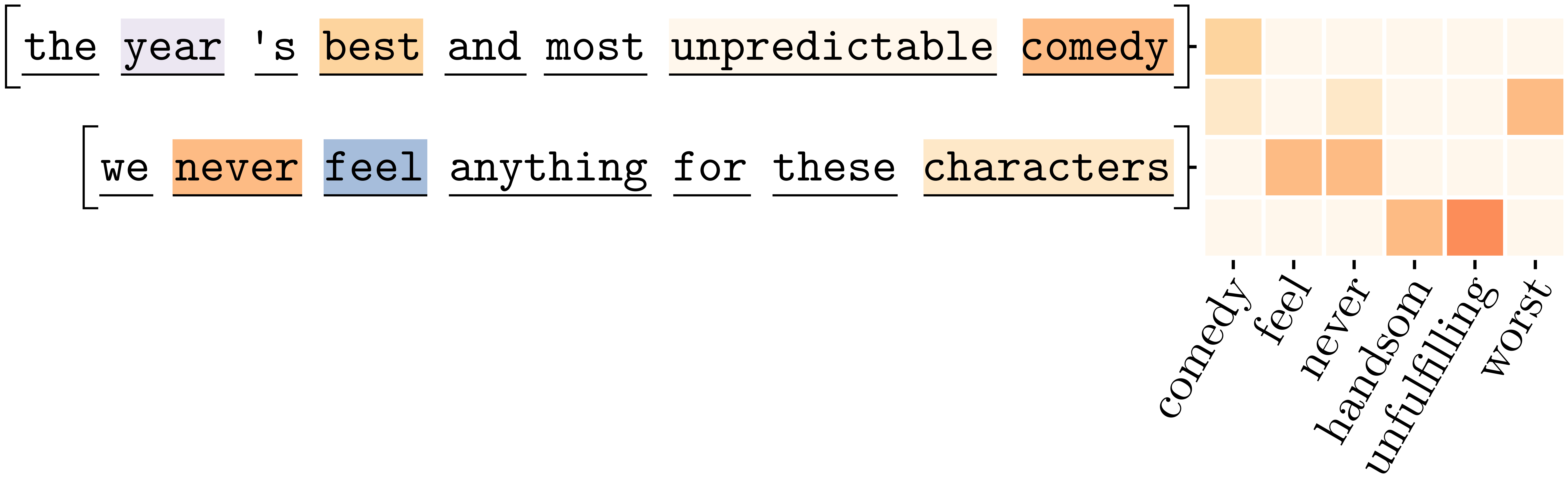
Global Explanation



How does the model relate words to each other?

Ensamble

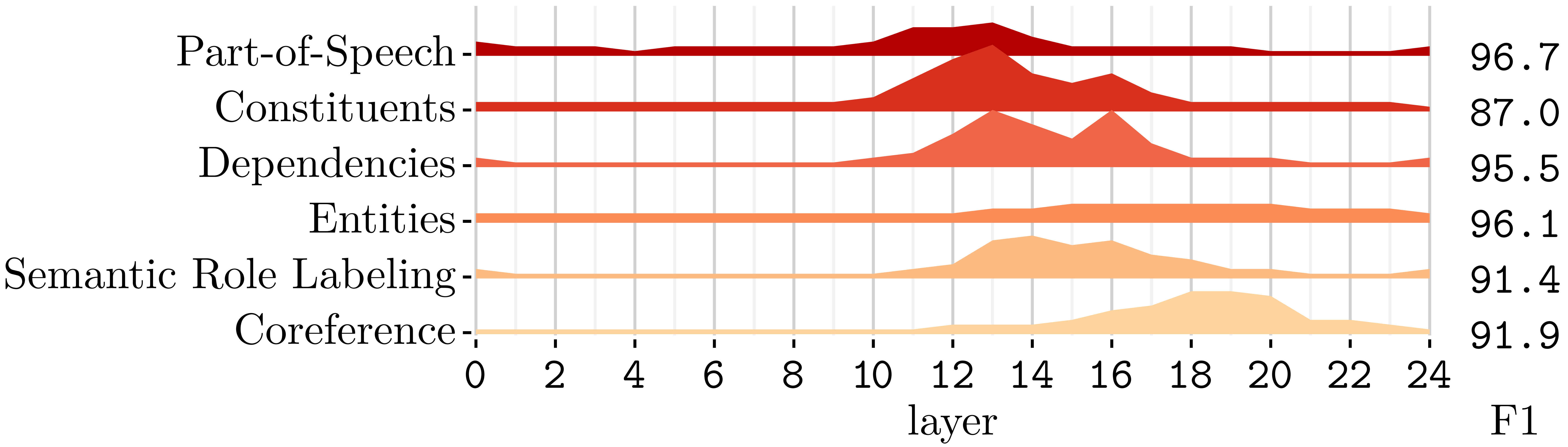
Global Explanation



What examples are representative of the model?

Linguistic information

Global Explanation



What linguistic information does the model use?

Rules

Global Explanation

		$p(y\|\mathbf{x})$	y	Flips
$\mathbf{x}$	`the year 's best and most unpredictable comedy`	0.91	1	-
$\tilde{\mathbf{x}}$	`the best and most unpredictable comedy this year`	0.13	-	-
rule	`DET year 's` $\rightarrow$ `this year`	-	-	1%
$\mathbf{x}$	`we never feel anything for these characters`	0.95	0	-
$\tilde{\mathbf{x}}$	`we never empatize for these characters`	0.11	-	-
rule	`feel` $\rightarrow$ `empatize`	-	-	4%

Which general rules can summarize an aspect of the model?

Takeaways

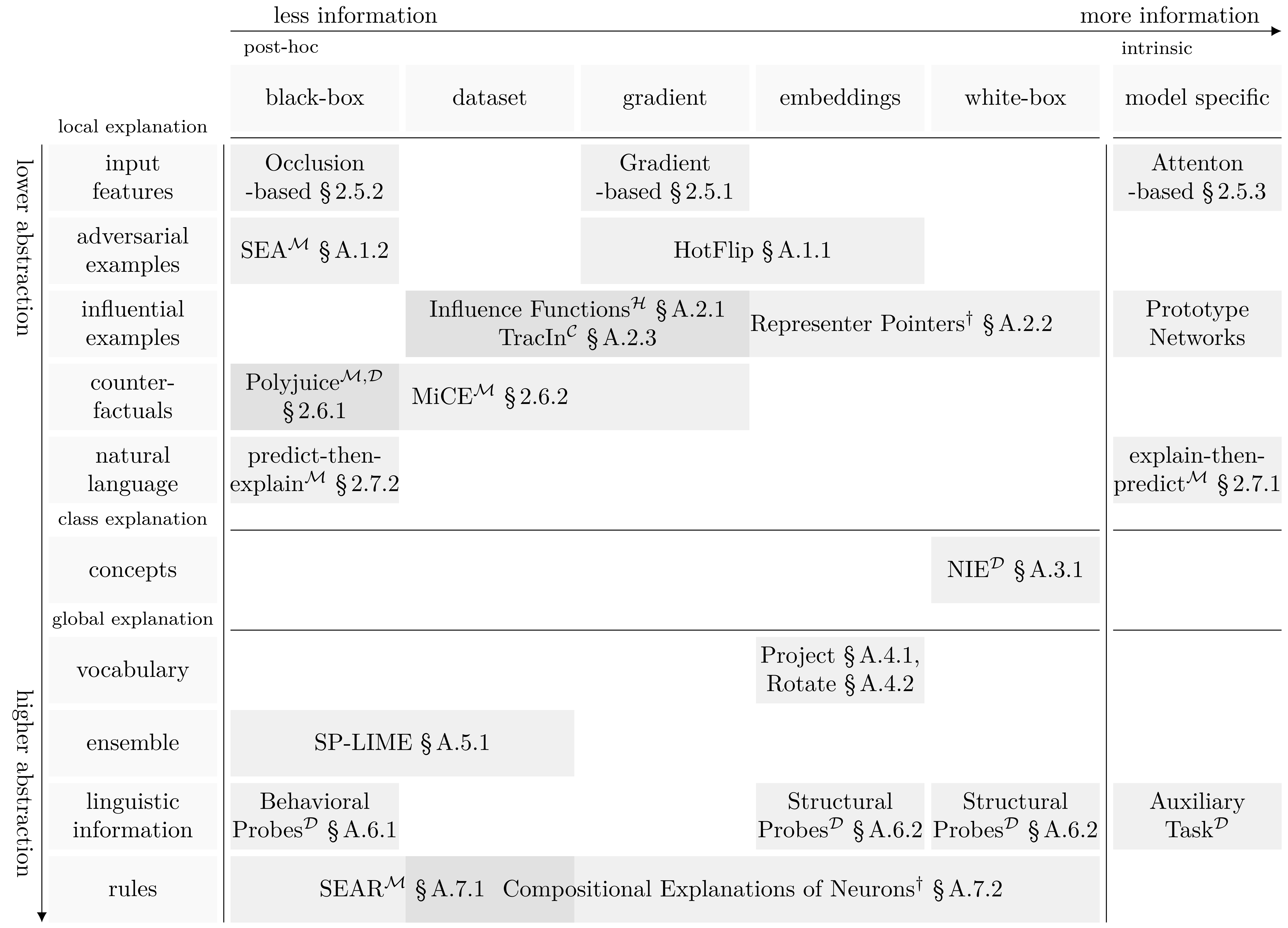
less information → more information

	post-hoc: black-box	dataset	gradient	embeddings	white-box	intrinsic: model specific
local explanation						
input features	Occlusion -based § 2.5.2		Gradient -based § 2.5.1			Attenton -based § 2.5.3
adversarial examples	SEA$^{\mathcal{M}}$ § A.1.2		HotFlip § A.1.1			
influential examples		Influence Functions$^{\mathcal{H}}$ § A.2.1 TracIn$^{\mathcal{C}}$ § A.2.3		Representer Pointers† § A.2.2		Prototype Networks
counter-factuals	Polyjuice$^{\mathcal{M},\mathcal{D}}$ § 2.6.1	MiCE$^{\mathcal{M}}$ § 2.6.2				
natural language	predict-then-explain$^{\mathcal{M}}$ § 2.7.2					explain-then-predict$^{\mathcal{M}}$ § 2.7.1
class explanation						
concepts					NIE$^{\mathcal{D}}$ § A.3.1	
global explanation						
vocabulary				Project § A.4.1, Rotate § A.4.2		
ensemble	SP-LIME § A.5.1					
linguistic information	Behavioral Probes$^{\mathcal{D}}$ § A.6.1			Structural Probes$^{\mathcal{D}}$ § A.6.2	Structural Probes$^{\mathcal{D}}$ § A.6.2	Auxiliary Task$^{\mathcal{D}}$
rules	SEAR$^{\mathcal{M}}$ § A.7.1	Compositional Explanations of Neurons† § A.7.2				

lower abstraction → higher abstraction

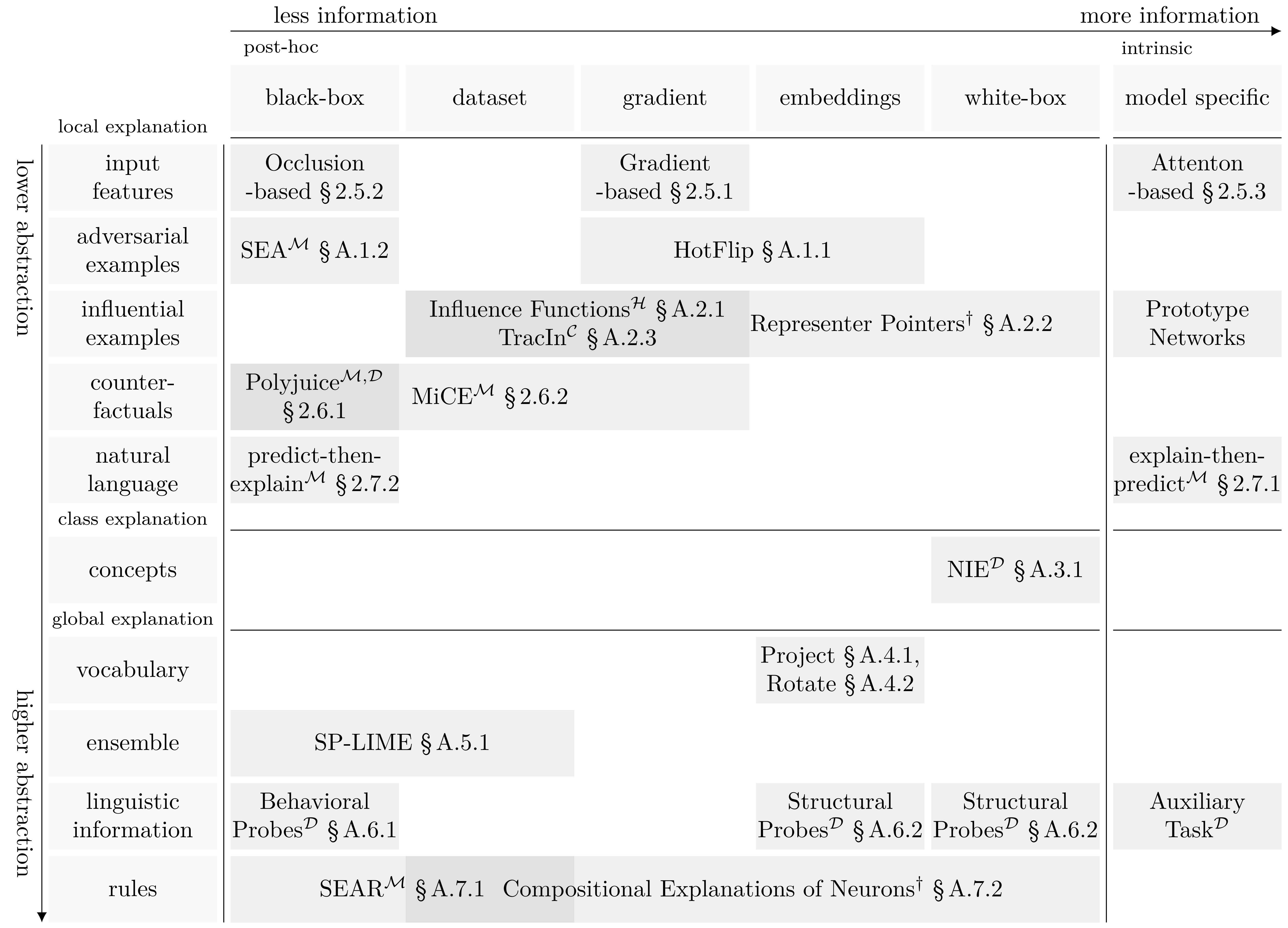
Takeaways

- Most methods are not evaluated well, and there have been little improvement.



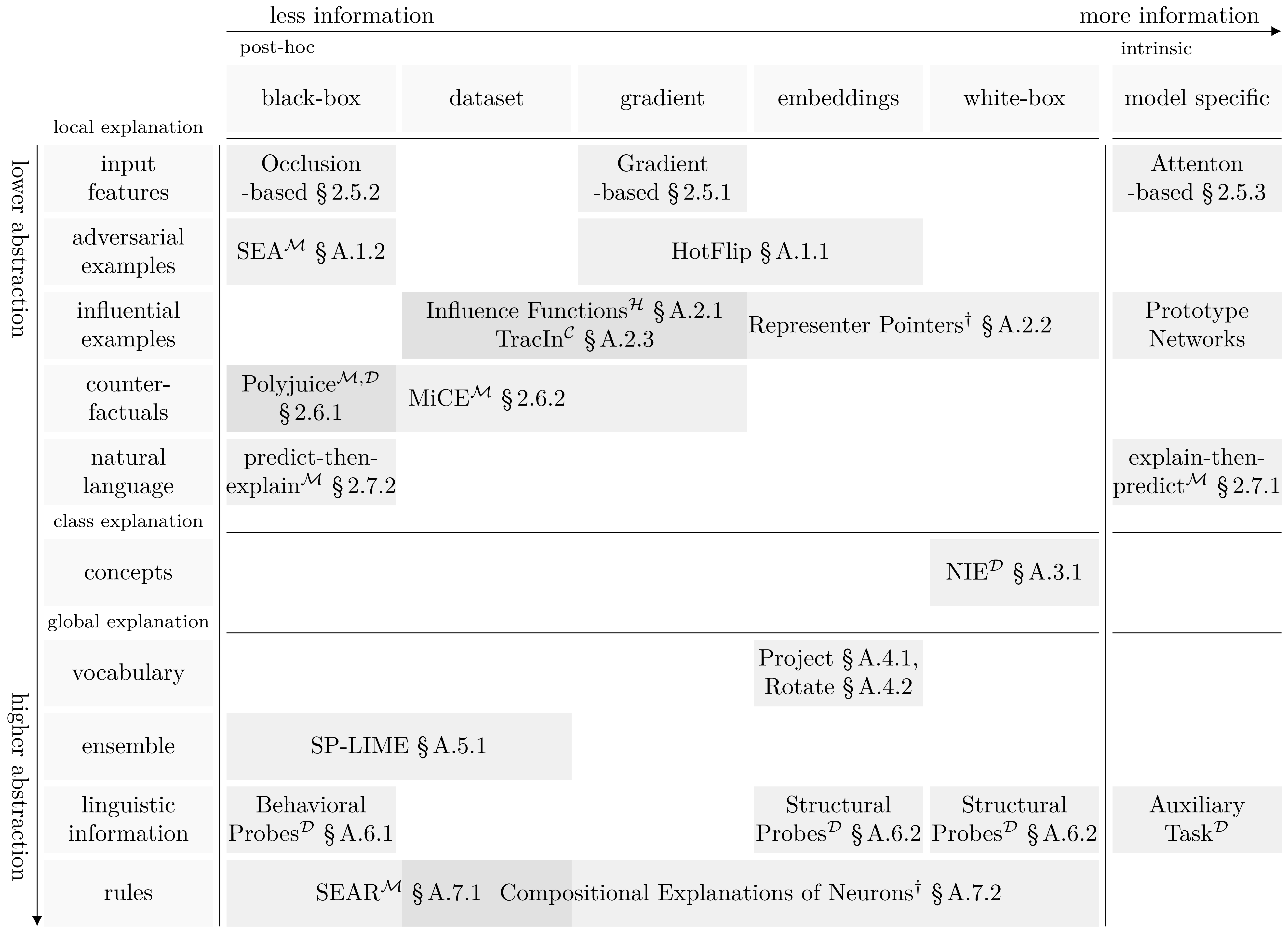
Takeaways

- Most methods are not evaluated well, and there have been little improvement.
- *Class explanation* methods is lacking, especially compared to computer vision.



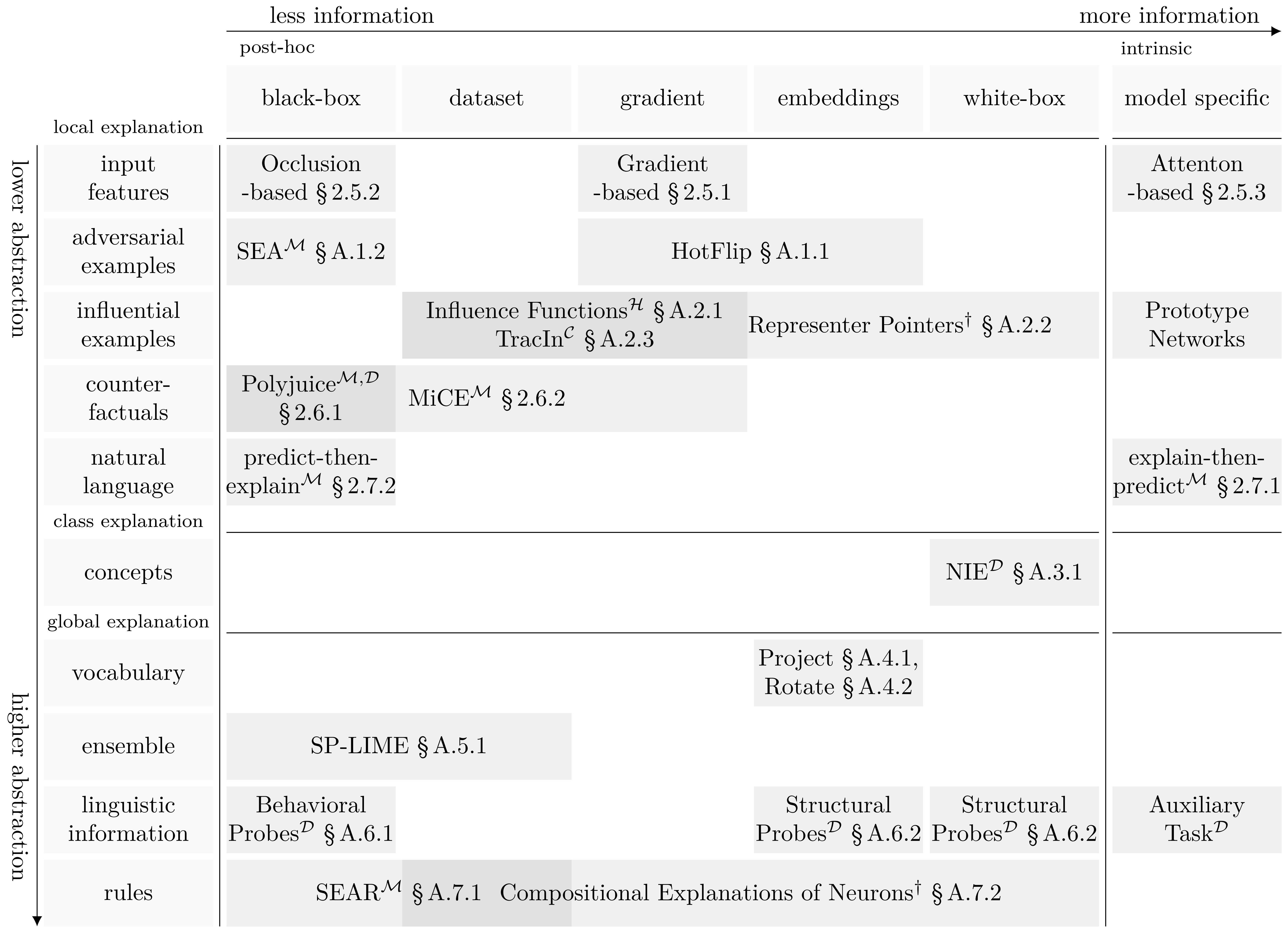
Takeaways

- Most methods are not evaluated well, and there have been little improvement.
- *Class explanation* methods is lacking, especially compared to computer vision.
- There is new work in computer vision that bridges the gap between *post-hoc* and *intrinsic*. Which is have not been adopted.



Takeaways

- Most methods are not evaluated well, and there have been little improvement.
- *Class explanation* methods is lacking, especially compared to computer vision.
- There is new work in computer vision that bridges the gap between *post-hoc* and *intrinsic*. Which is have not been adopted.
- Large Pre-trained models, like GPT-2 and T5, have enabled great progress in creating fluent explanations.



Recursive-ROAR

Desirables

a) The method does not assume a known true explanation.
b) The method measures faithfulness of an explanation w.r.t. a specific model instance and single observation. For example, it is not a proxy-model that is measured.
c) The method uses only the original dataset, e.g. does not introduce spurious correlations.
d) The method only uses inputs and intermediate representations that are in-distribution w.r.t. the model.
e) The method is computationally cheap by not training/fine-tuning repeatedly and only computes explanations of the test dataset.
f) The method can be applied to any classification task.
g) The method can be applied to any importance measure.

Recursive ROAR: satisfies (a), (c), (d), (f), and (g).

Leaking target variable

Thought experiment

a) Say “awful” is a strong indicator of negative sentiment.
b) Recursive ROAR will remove “awful” from every negative sentiment observation.
c) “awful” is now a perfect predictor of positive sentiment.
   e.g. “I have an awful strong crush on this actor”

We want an importance measure for the correct label, as removing the tokens that are relevant for making a wrong prediction, would help the performance of the model.

Leaking target variable

- $\{x_1, x_2, x_3, x_4\}$ are relevant features, but mutually redundant. All other features are irrelevant to the target value.

- z, η, ϵ are sampled for each observation. r_i, s_i are sampled once. A standard normal distribution is used.

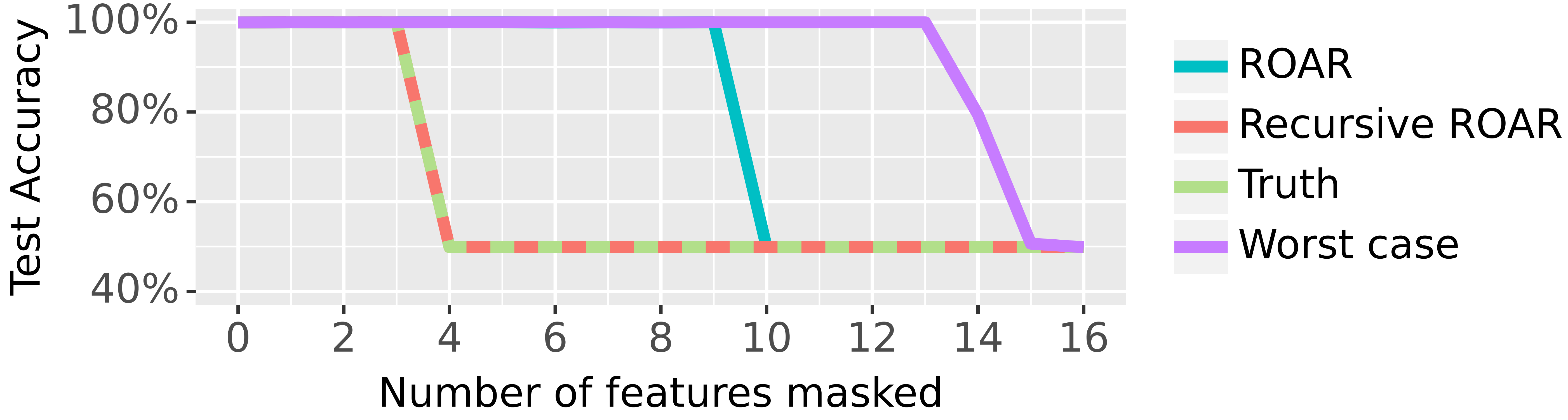
- The explanation is the weights of a logistic regression.

$$\mathbf{x} = \frac{\mathbf{a}z}{10} + \mathbf{d}\eta + \frac{\epsilon}{10}, \quad y = \begin{cases} 1 & z > 0 \\ 0 & z \leq 0 \end{cases}$$

$$a = [r_1, r_2, r_3, r_4, 0,0,0,0,0,0,0,0,0,0,0,0]$$

$$d = [s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8, s_9, s_{10}, s_{11}, s_{12}, s_{13}, s_{14}, s_{15}, s_{16}]$$

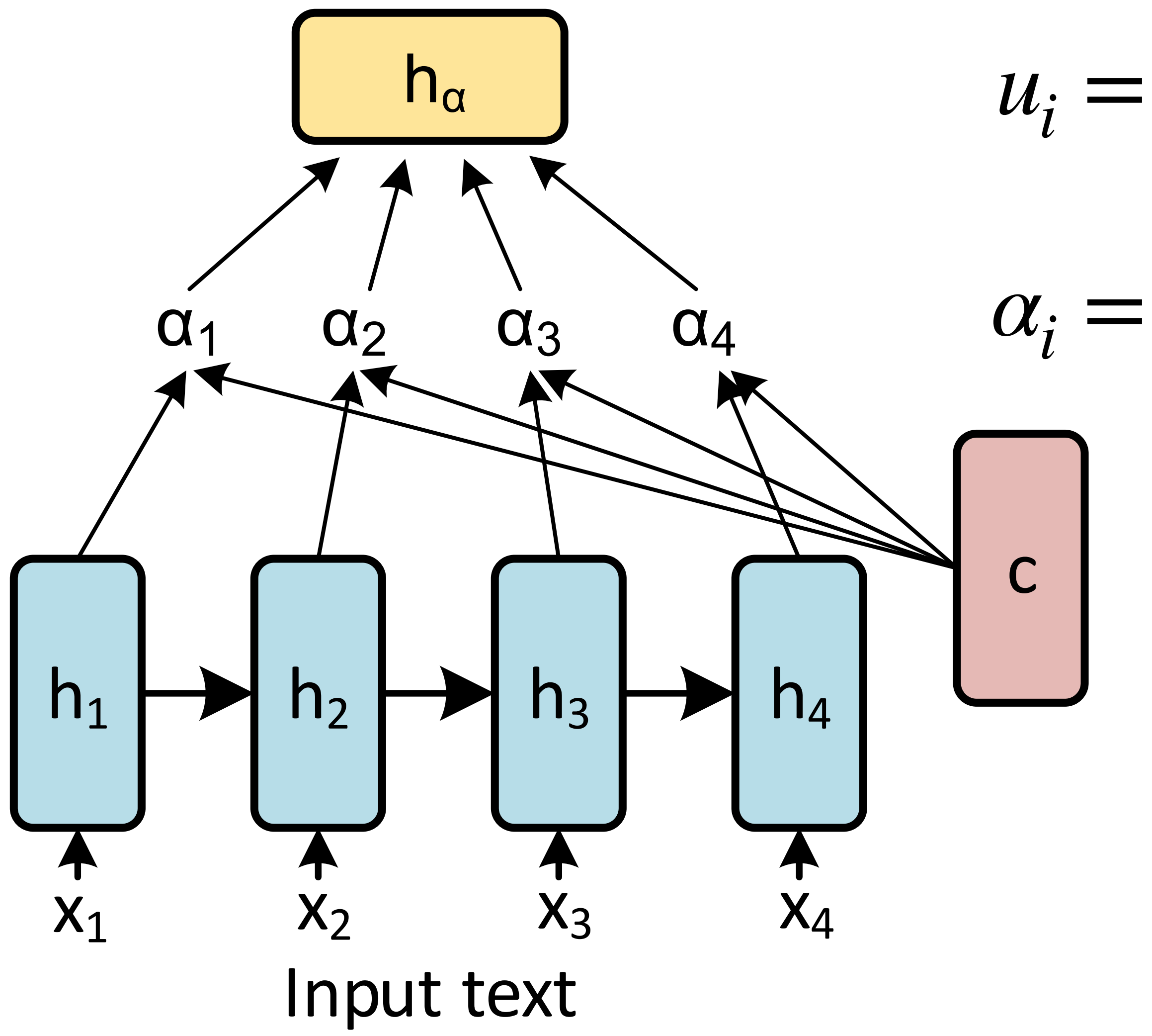
$$x = \left[\frac{r_1 z}{10} + s_1\eta + \frac{\epsilon}{10}, \cdots, \frac{r_4 z}{10} + s_4\eta + \frac{\epsilon}{10}, s_5\eta + \frac{\epsilon}{10}, \cdots, s_{16}\eta + \frac{\epsilon}{10}\right]$$



Attention Models

single sequence to class

Tasks: SST, IMDB, Anemia, Diabetes

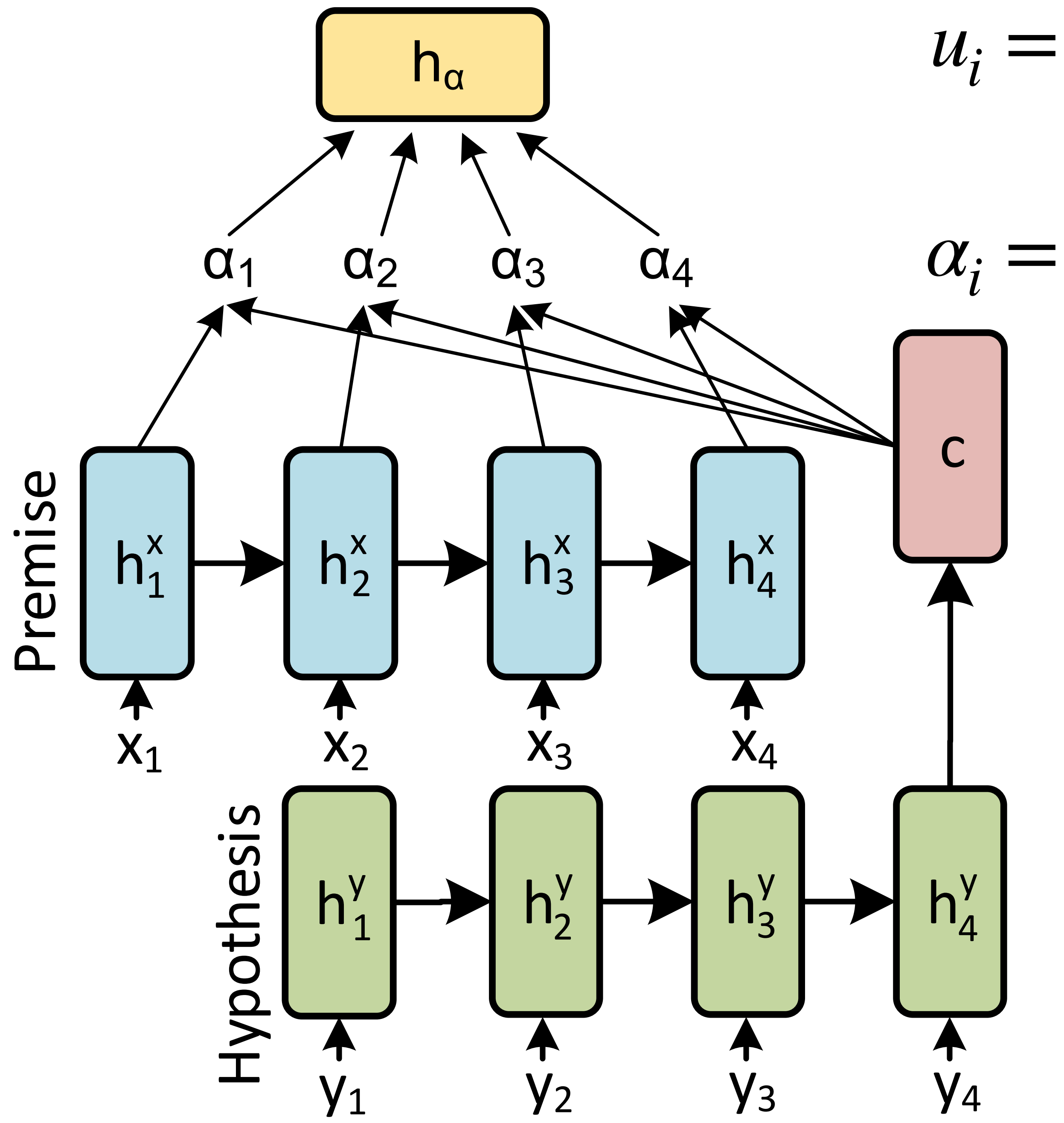


$$u_i = \tanh(W_1 h_i^x + b)$$

$$\alpha_i = \frac{\exp(u_i^T c)}{\sum_j \exp(u_j^T c)}$$

paired sequence to class

Tasks: SNLI, bAbI-1, bAbI-2, bAbI-3



$$u_i = \tanh(W_1 h_i^x + W_2 h_{T_2}^y)$$

$$\alpha_i = \frac{\exp(u_i^T c)}{\sum_j \exp(u_j^T c)}$$

[1] Vashishth et al, arXiv 2019, "Attention Interpretability Across NLP Tasks"
[2] Jain, ACL 2019, "Attention is not Explanation".

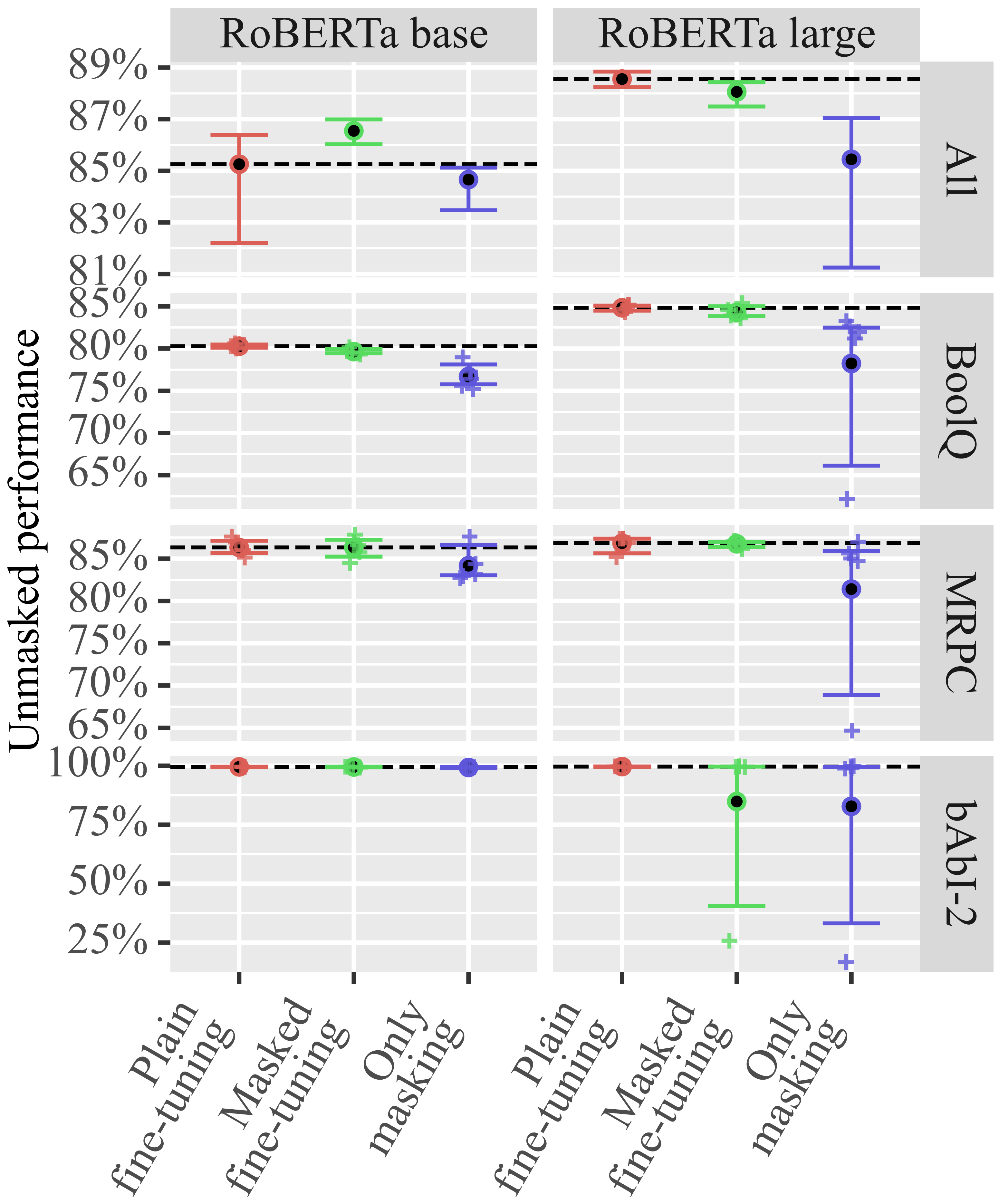
Papers on the faithfulness of attention

Paper	Compare with other importance measure	Test if mutated attention can yield same prediction	Test if learned adversarial attention can yield same prediciton.
Attention is not explanation (ACL 2019)	x	x	
Attention is not not explanation (EMNLP 2019)			x
Attention interpretability Across NLP Tasks (ArXiv 2019, ICLR 2020 Reject)		x	x
Is Attention Interpretable (ACL 2019)		x	
Learning to Deceive with Attention-Based Explanations (ACL 2020)			x
Is Sparse Attention more Interpretable (ACL 2021)	x	x	x
	Criticism: Other methods are not ground-truths.	**Criticism:** Mutating the attention causes out-out-distribution issues.	**Criticism:** Learning a different models says nothing about the original model.

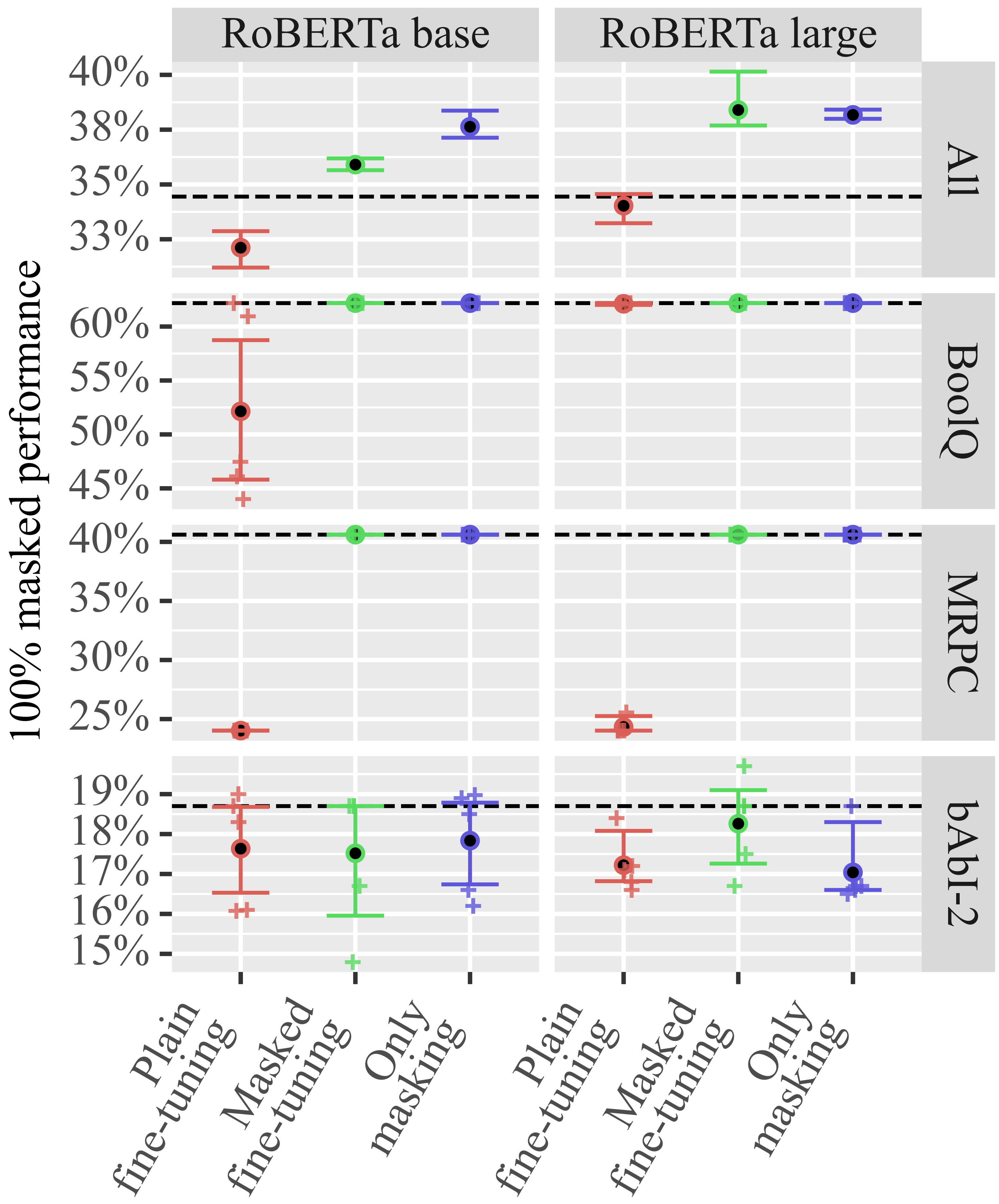
FMM

No performance issues

0% masked performance

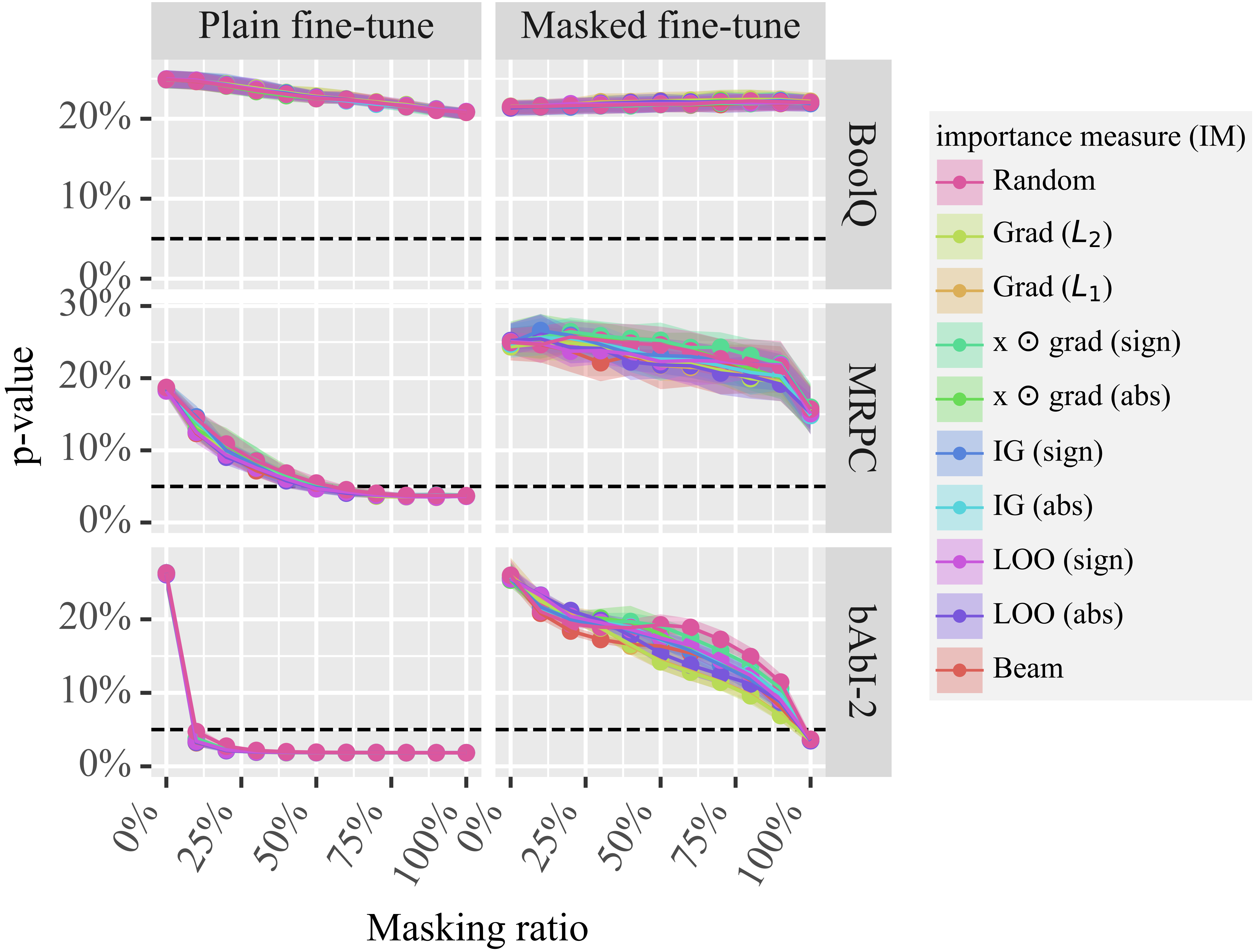


100% masked performance

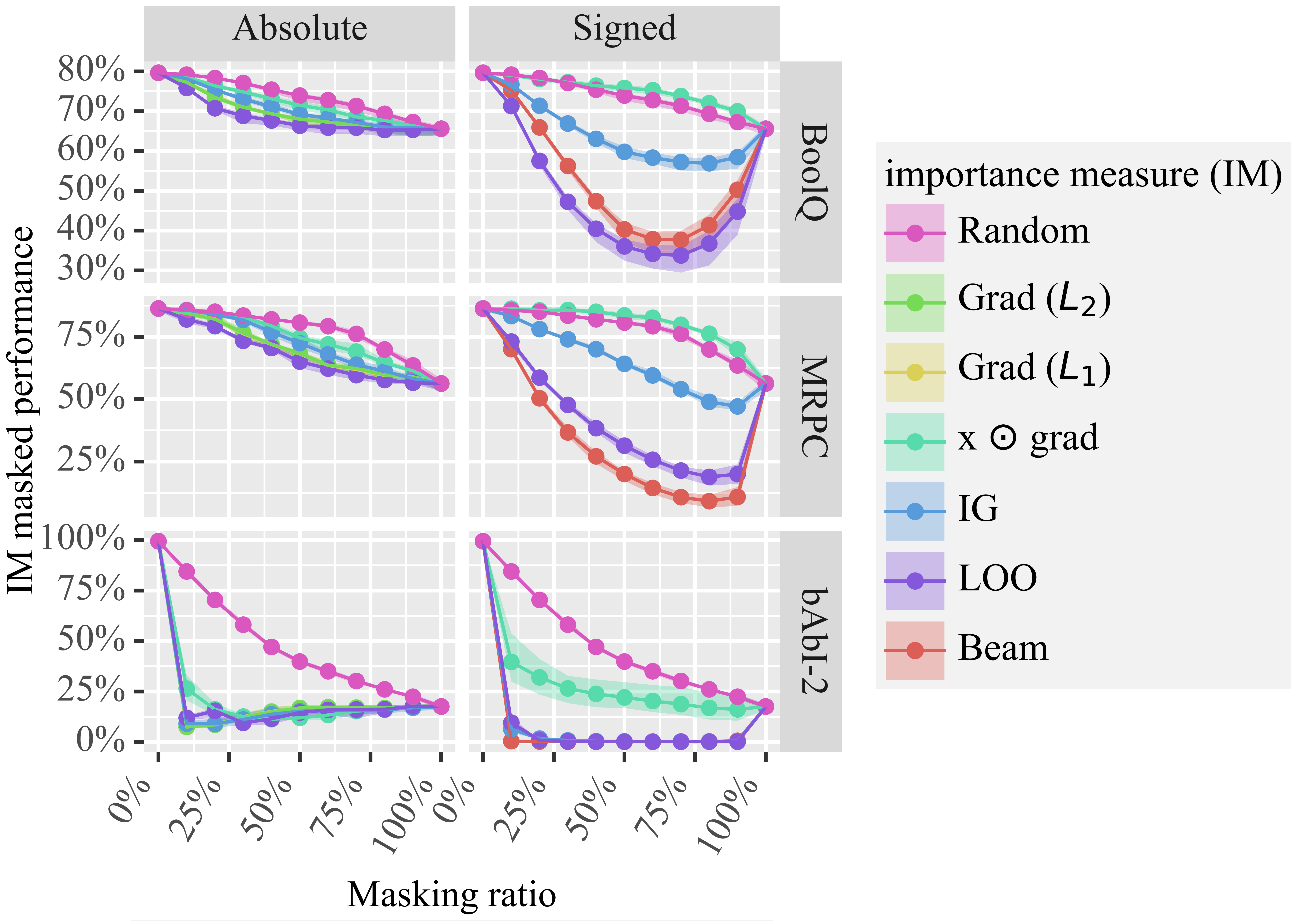


In-distribution testing

- Because random masking is different form targeted masking,
  each explanation need to be tested.
- Often out-of-distribution issues with plain fine-tuning.
- No out-of-distribution issues with masked fine-tuning.



Faithfulness



Masked CLM

Sequential output

Requirements are: 1) performance metric and 2) importance measure / ranking.

1. **Performance Measure:** ROUGE, BLEU, Levenstein.

2. **Importance measure:** Leave-on-out, naive aggregation, optimization, etc.

Masked CLMs

Learn masking support during pre-training

Mask random tokens during pre-training with a next-token objective.

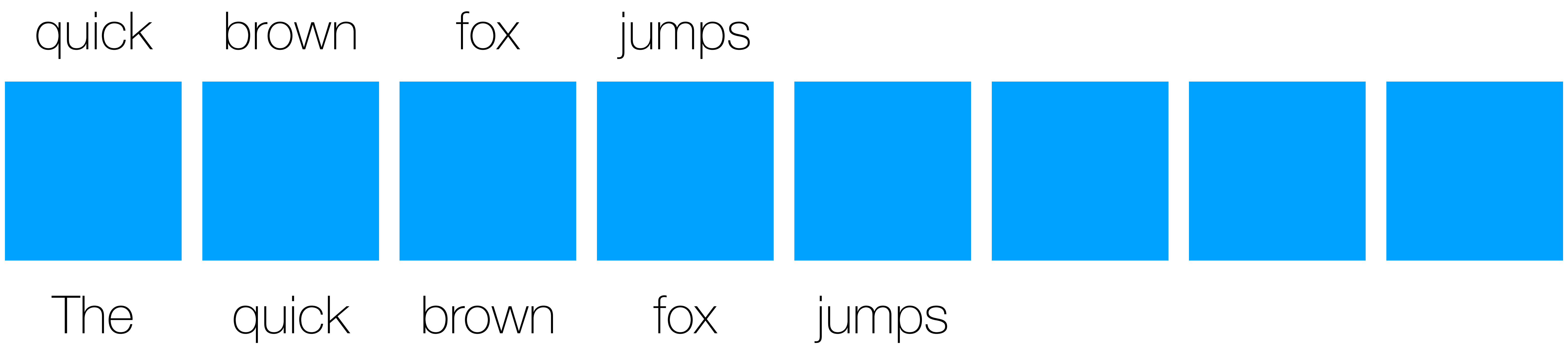
Masked CLMs

Learn masking support during pre-training

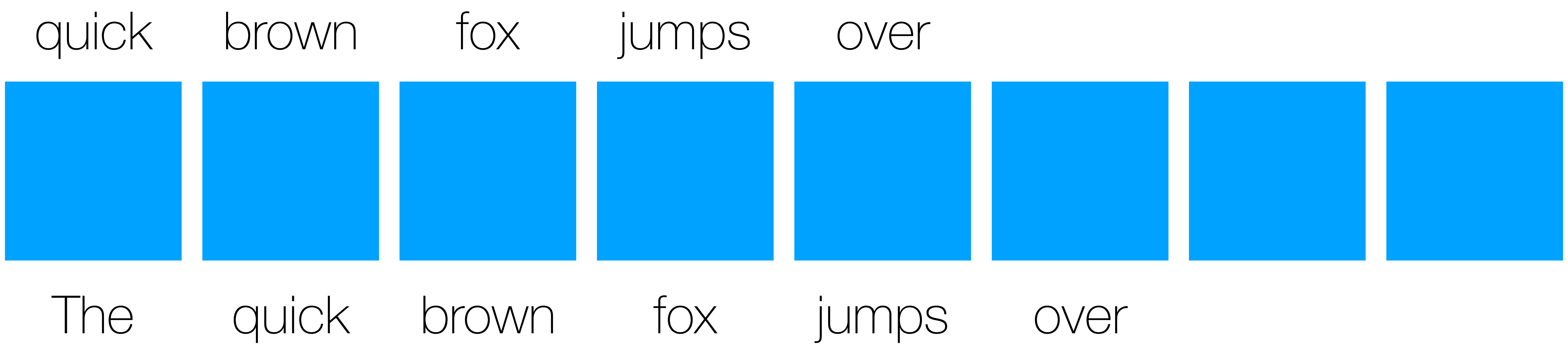
Mask random tokens during pre-training with a next-token objective.

1. An Faithfulness Measurable model.
2. Get highly faithful occlusion-based importance measure.

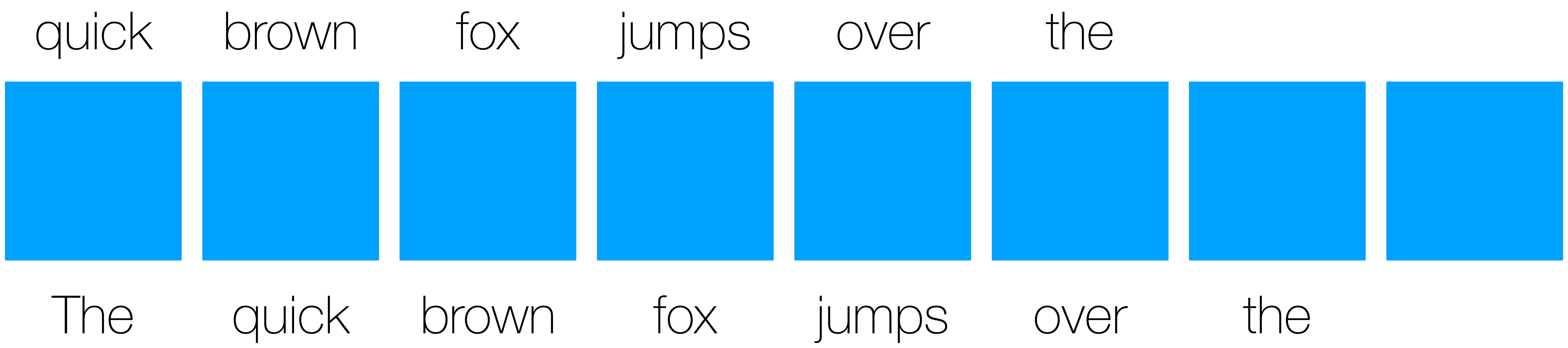
Masked CLMs



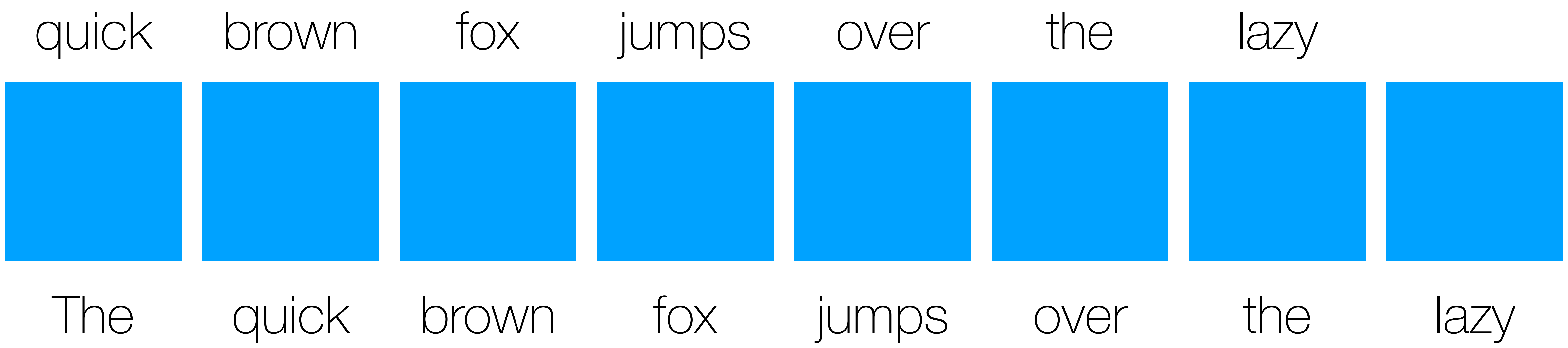
Masked CLMs



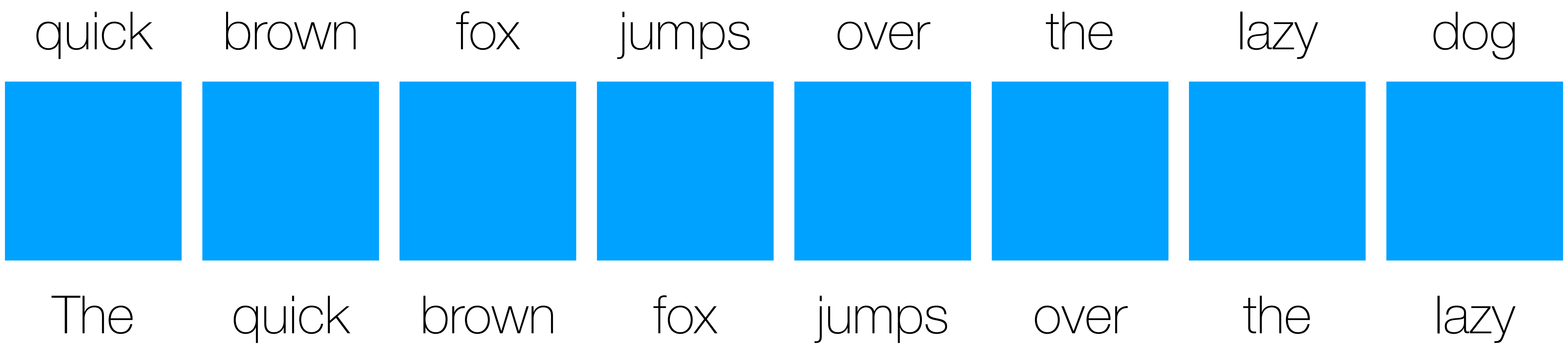
Masked CLMs


quick brown fox jumps over the
The quick brown fox jumps over the

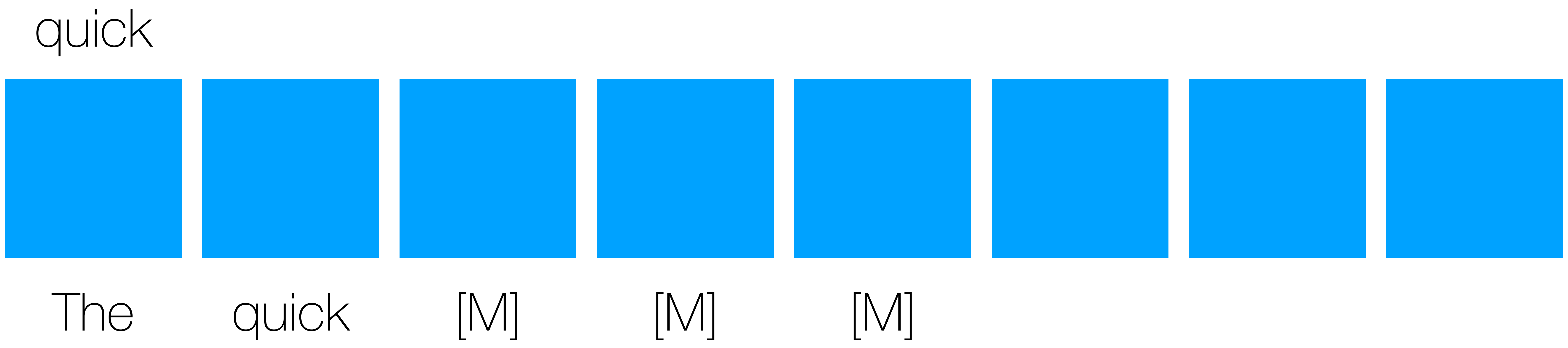
Masked CLMs


quick brown fox jumps over the lazy
The quick brown fox jumps over the lazy

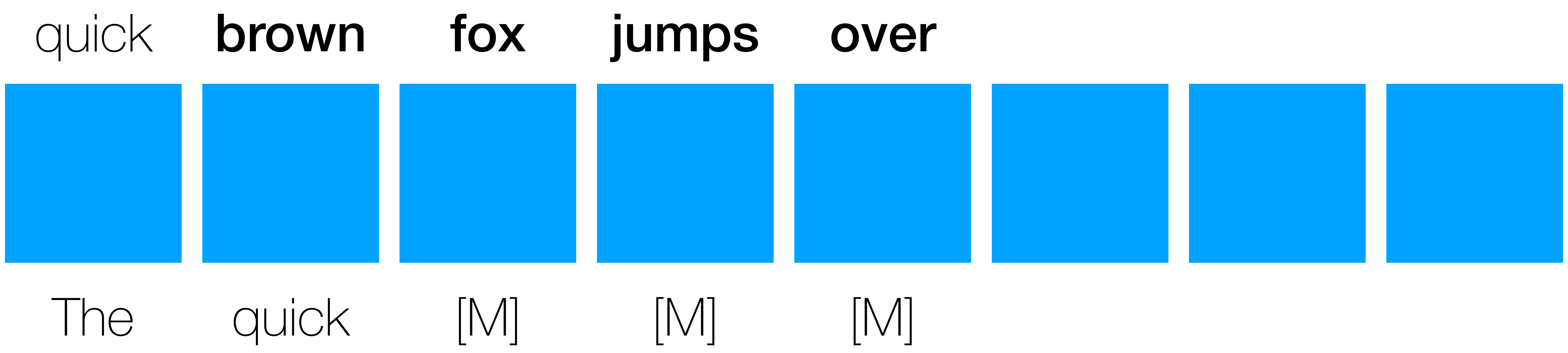
Masked CLMs


quick brown fox jumps over the lazy dog
The quick brown fox jumps over the lazy

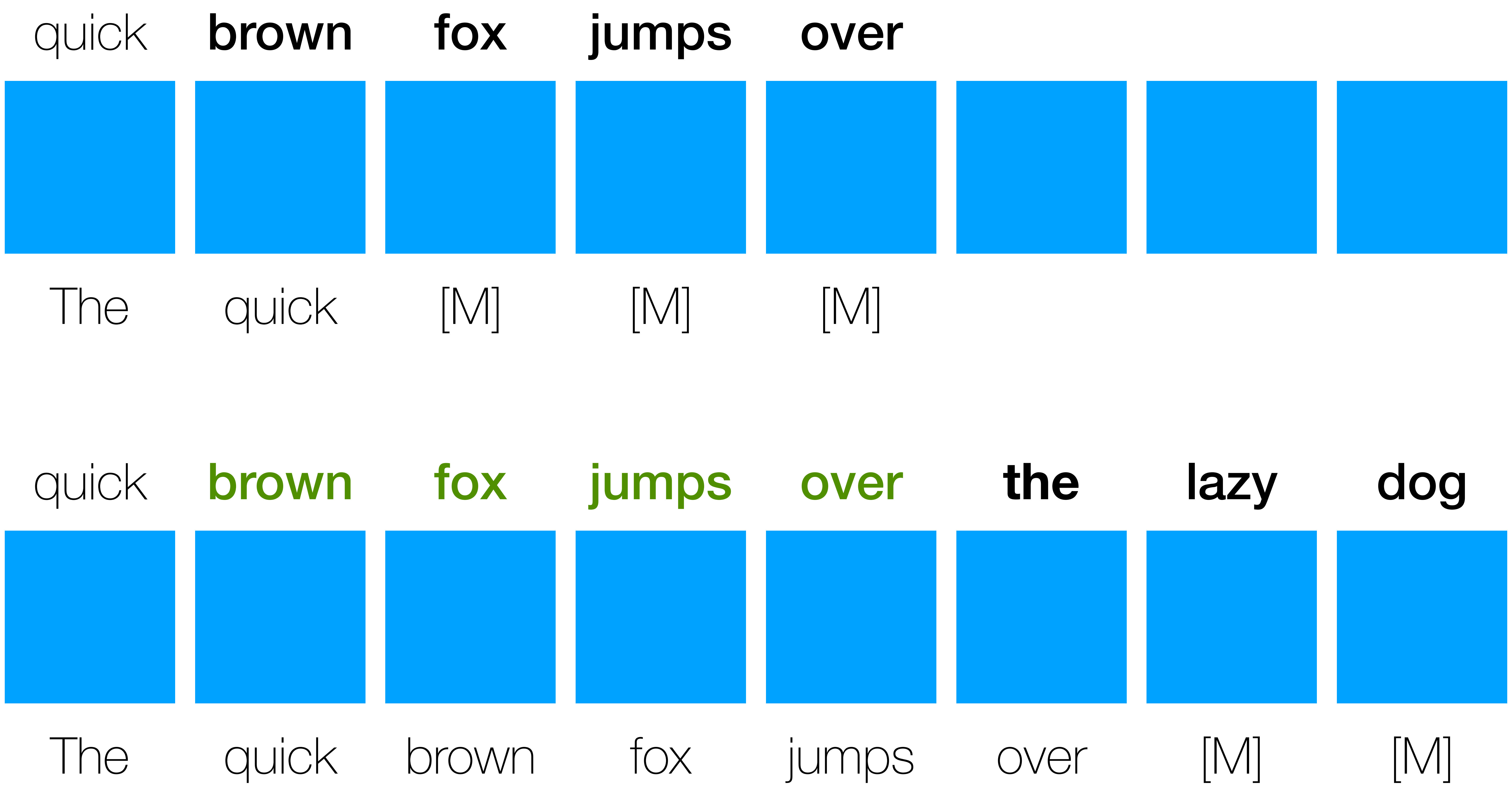
Masked CLMs



Masked CLMs



Masked CLMs


quick brown fox jumps over
The quick [M] [M] [M]
quick brown fox jumps over the lazy dog
The quick brown fox jumps over [M] [M]

Masked CLMs

Learn masking support during pre-training

Mask random tokens during pre-training with a next-token objective.

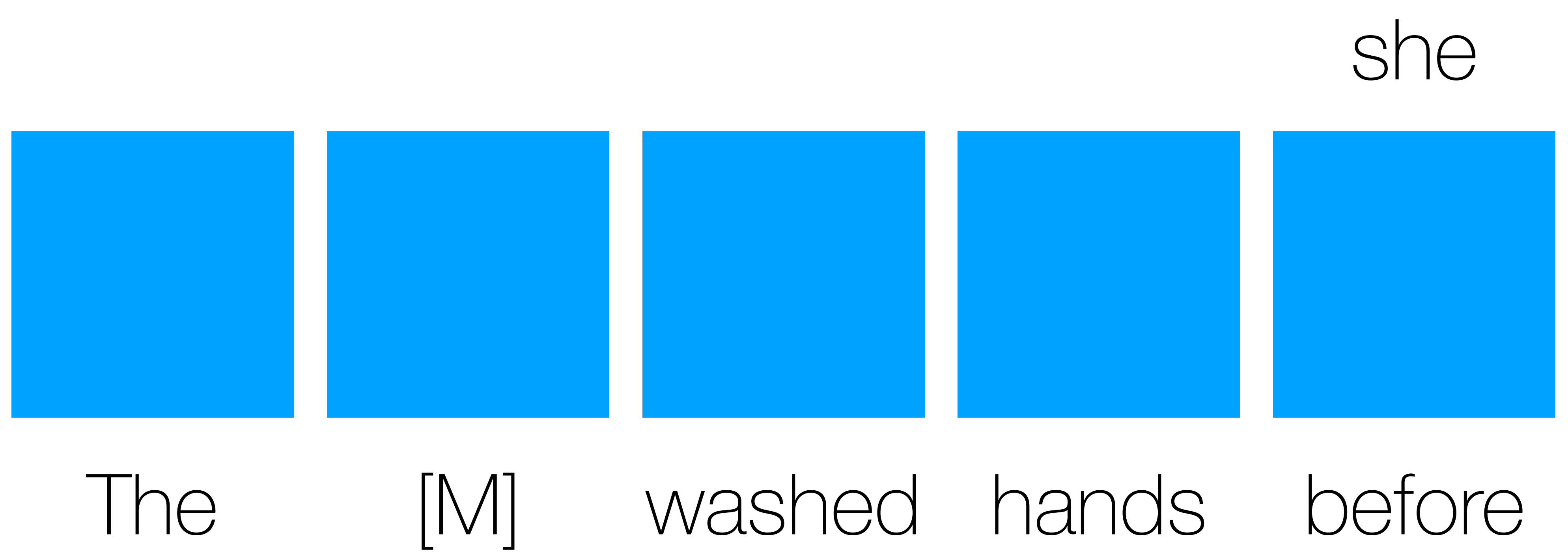
1. An Faithfulness Measurable model.
2. Get highly faithful occlusion-based importance measure.
3. Zero-cost parallel-token generation.

Masked CLMs

Learn masking support during pre-training

Mask random tokens during pre-training with a next-token objective.

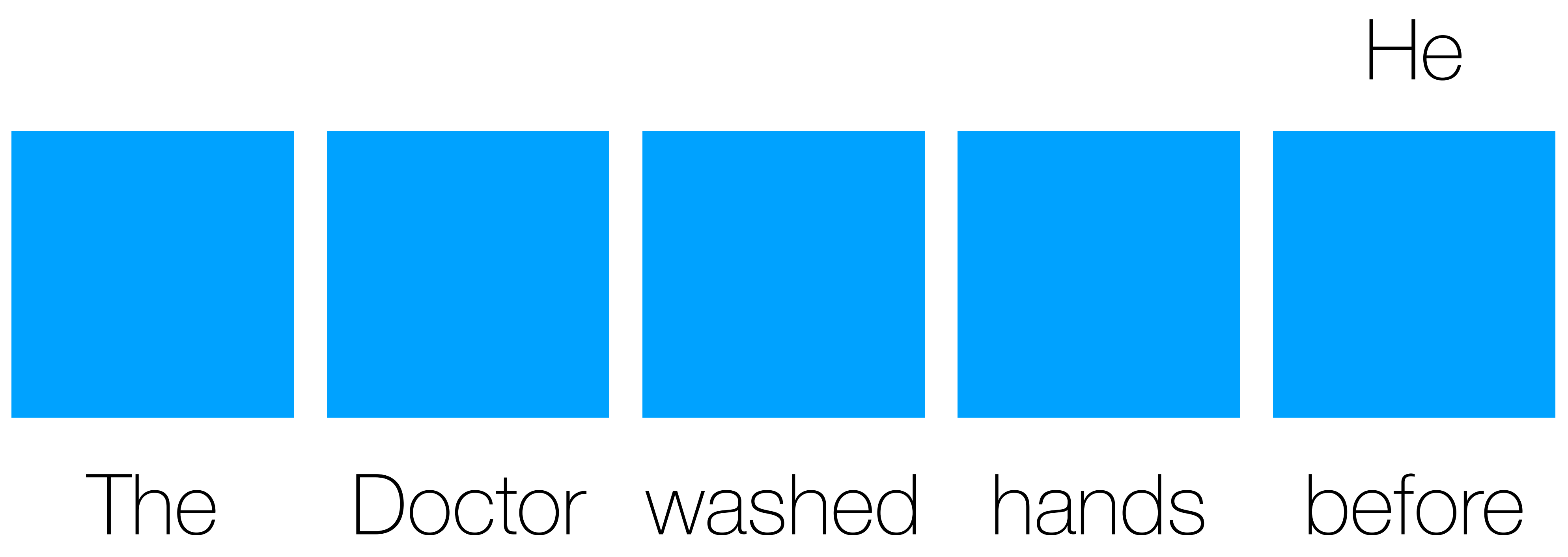
1. An Faithfulness Measurable model.
2. Get highly faithful occlusion-based importance measure.
3. Zero-cost parallel-token generation.
4. Many established techniques from MLM.



Masked CLMs

Learn masking support during pre-training
Mask random tokens during pre-training with a next-token objective.

1. An Faithfulness Measurable model.
2. Get highly faithful occlusion-based importance measure.
3. Zero-cost parallel-token generation.
4. Many established techniques from MLM.

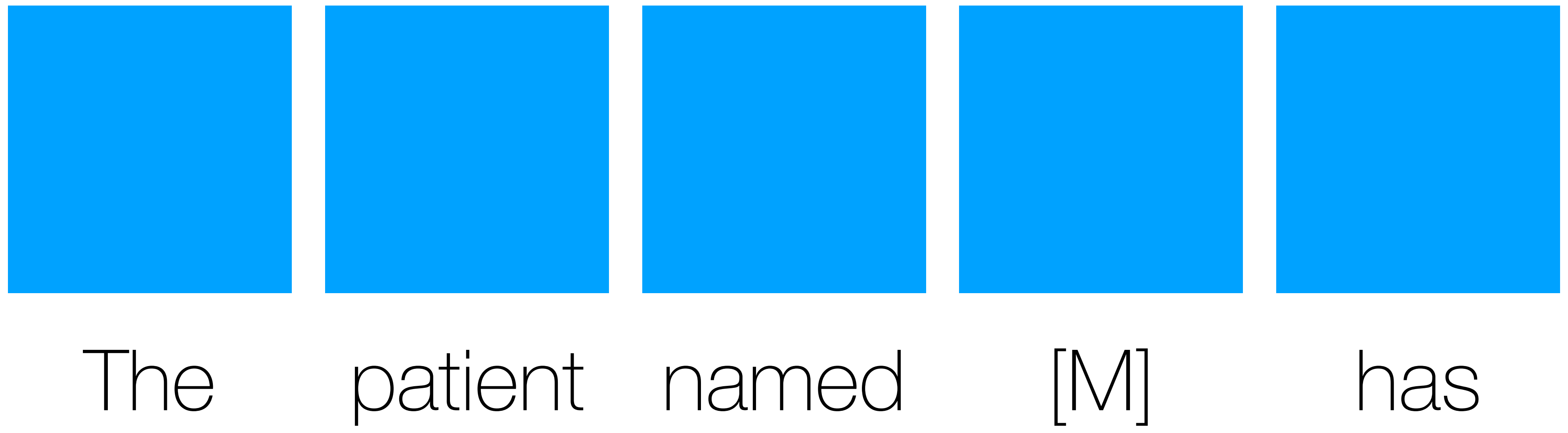


Masked CLMs

Learn masking support during pre-training

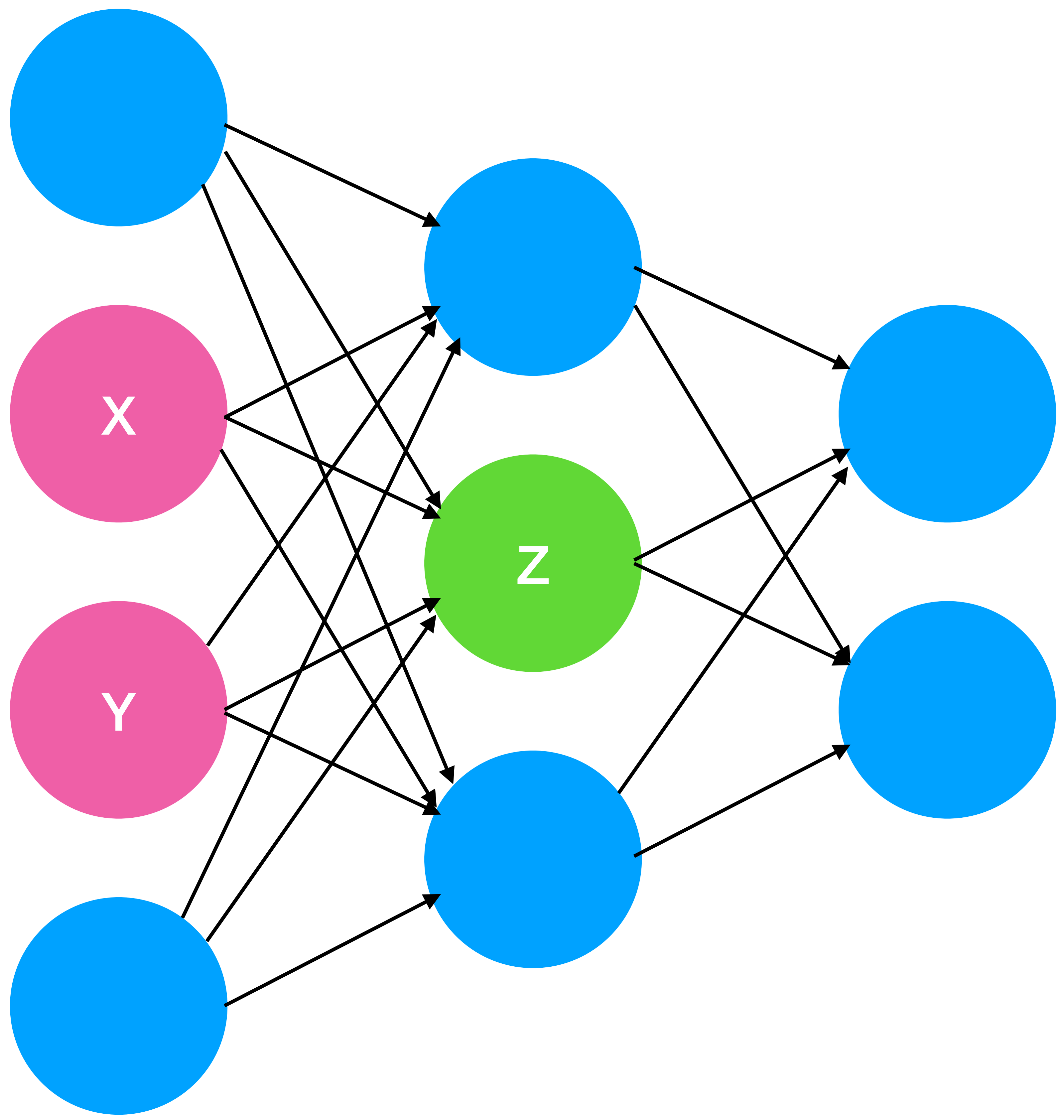
Mask random tokens during pre-training with a next-token objective.

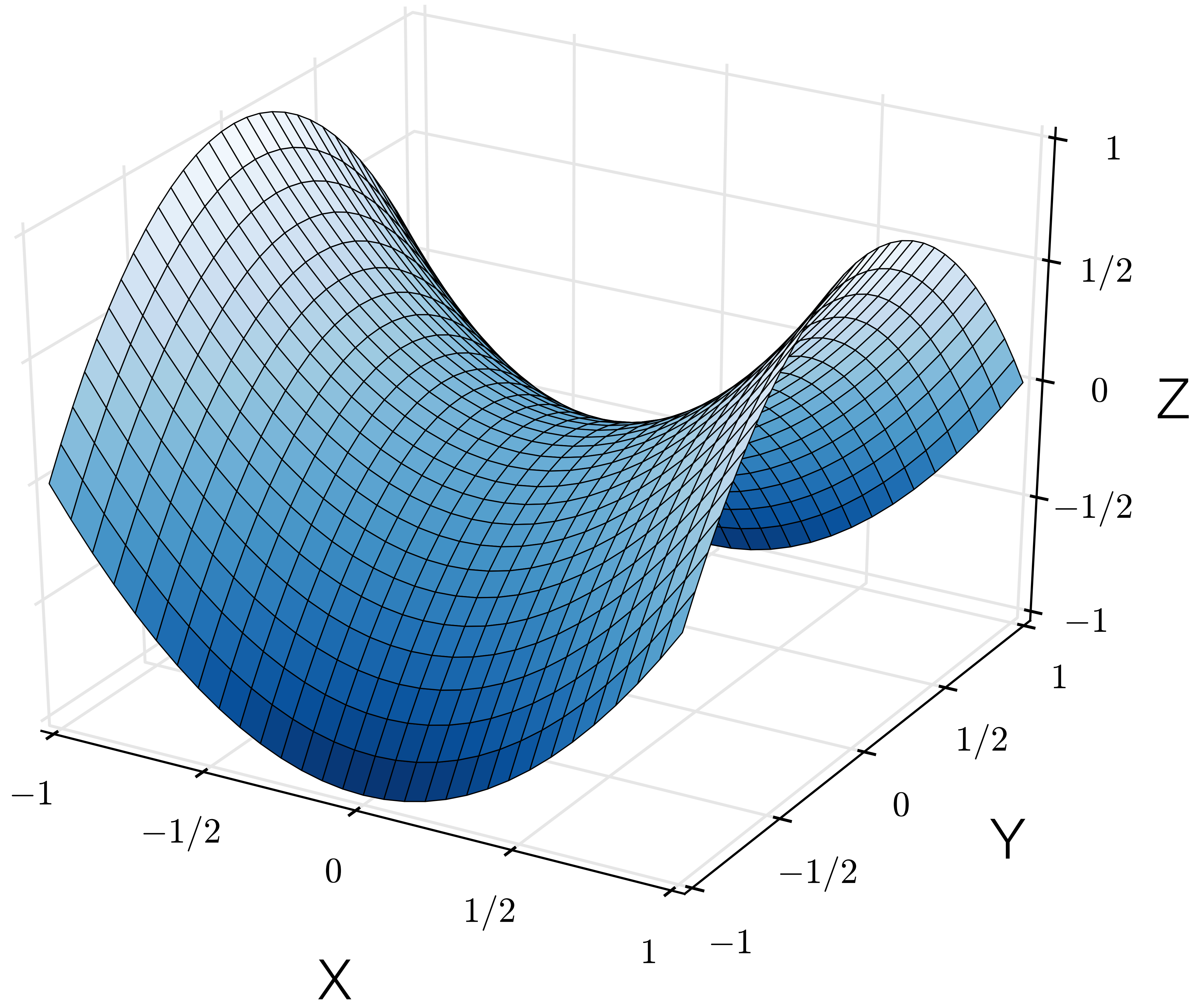
1. An Faithfulness Measurable model.
2. Get highly faithful occlusion-based importance measure.
3. Zero-cost parallel-token generation.
4. Many established techniques from MLM.
5. Standard for how to anonymize data.



MaSF

Manifolds

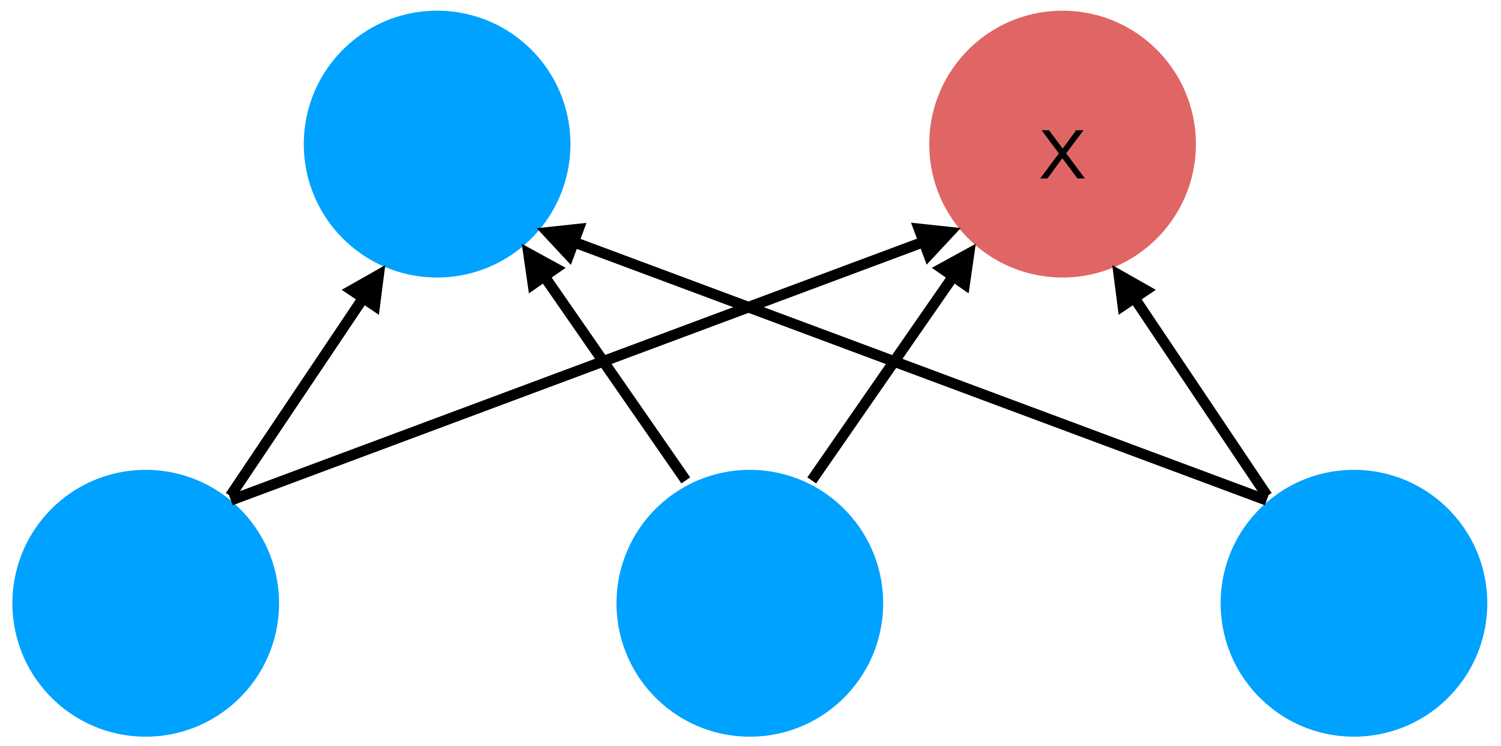




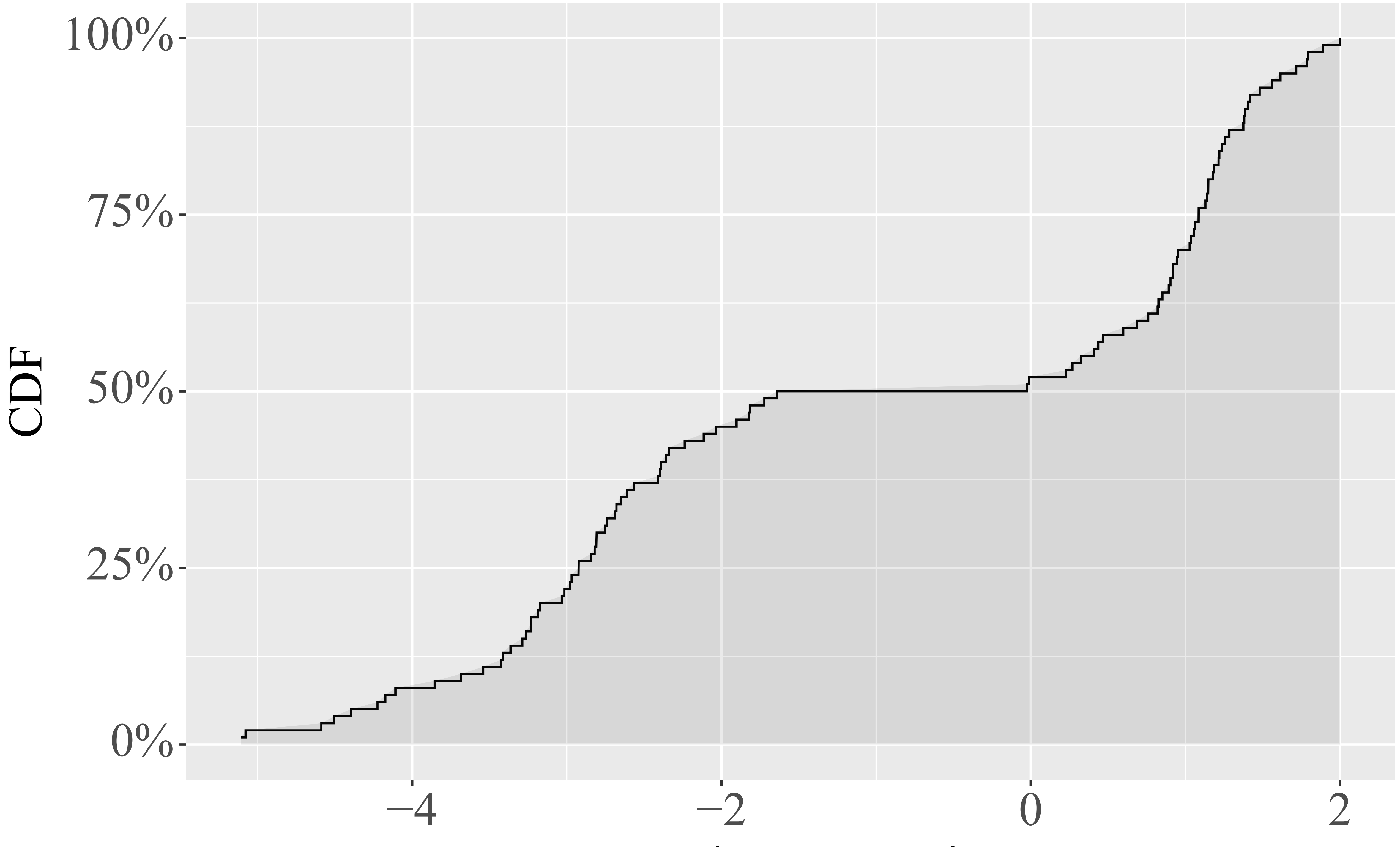
Desirables

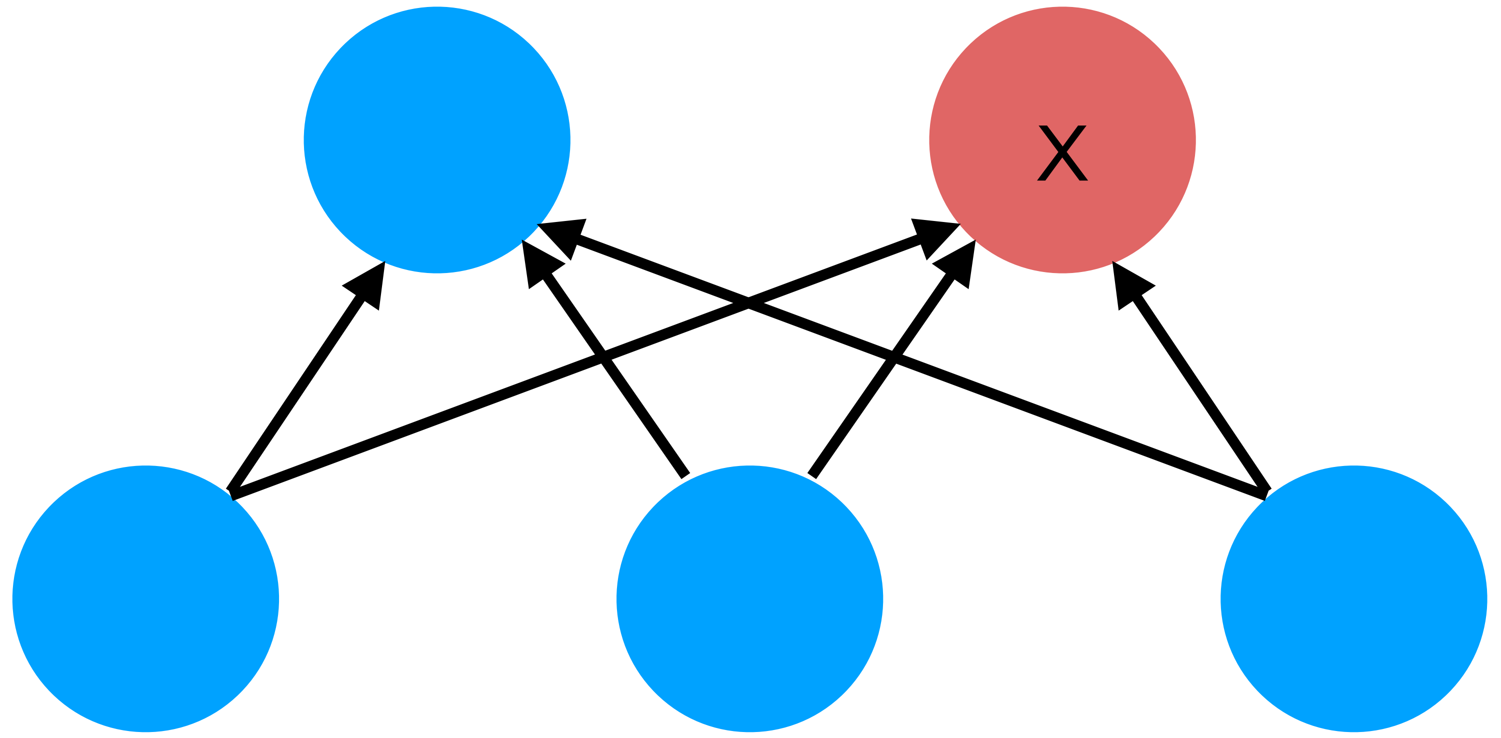
- Should assume little of the model’s internals. For example, do not assume internally normally distributed.
- Should only consider the model, not the input distribution (sensory anomaly detection).
- Should provide non-ambiguous metrics.

Empirical CDF

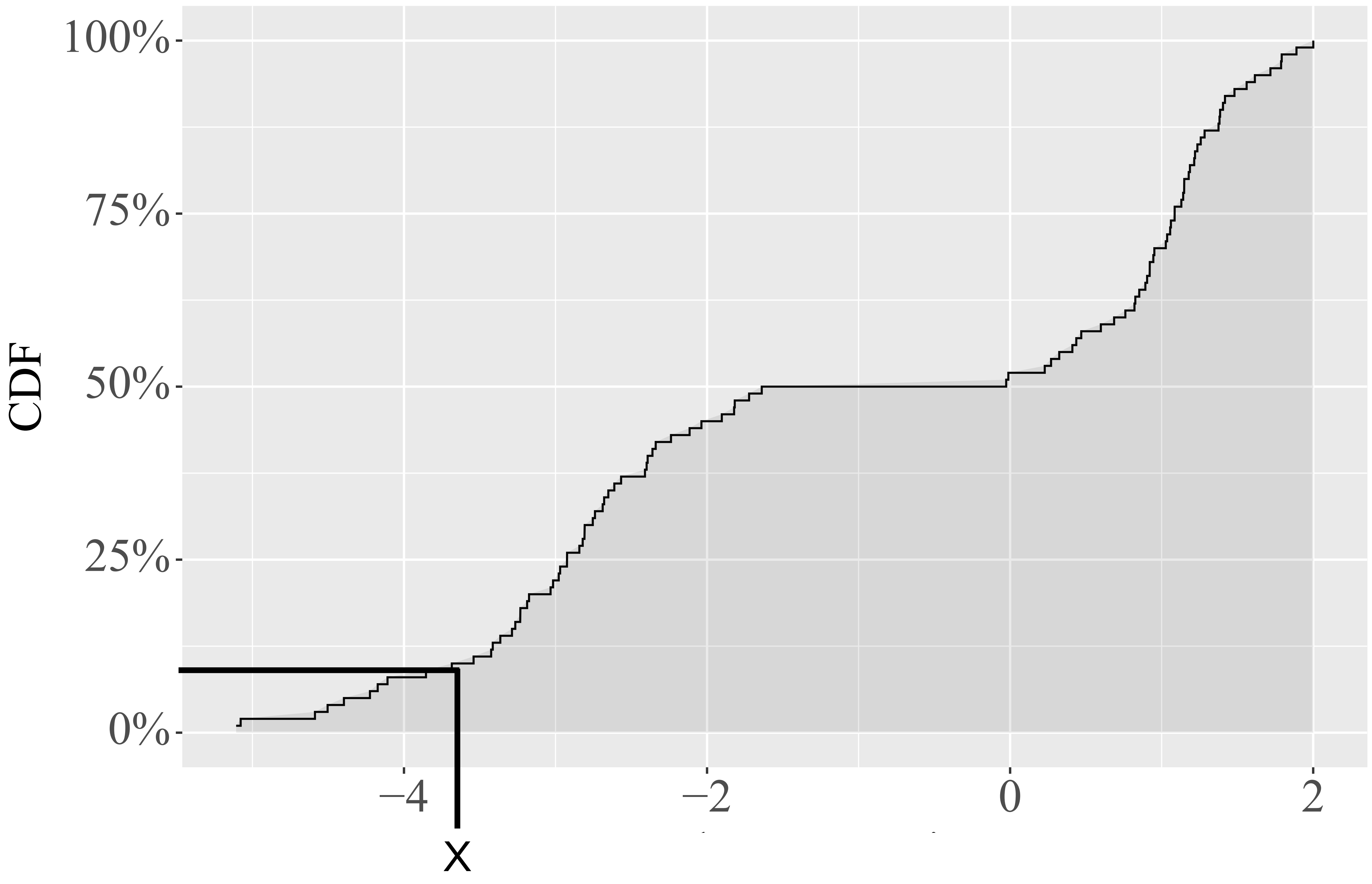


Empirical CDF





Empirical CDF



One-sided p-value

$$p = \mathbb{P}(X \leq x)$$
$$\approx \frac{1}{|D|} \sum_{v \in D} 1[v \leq x] \quad \text{where } D \sim X$$

Two-sided p-value

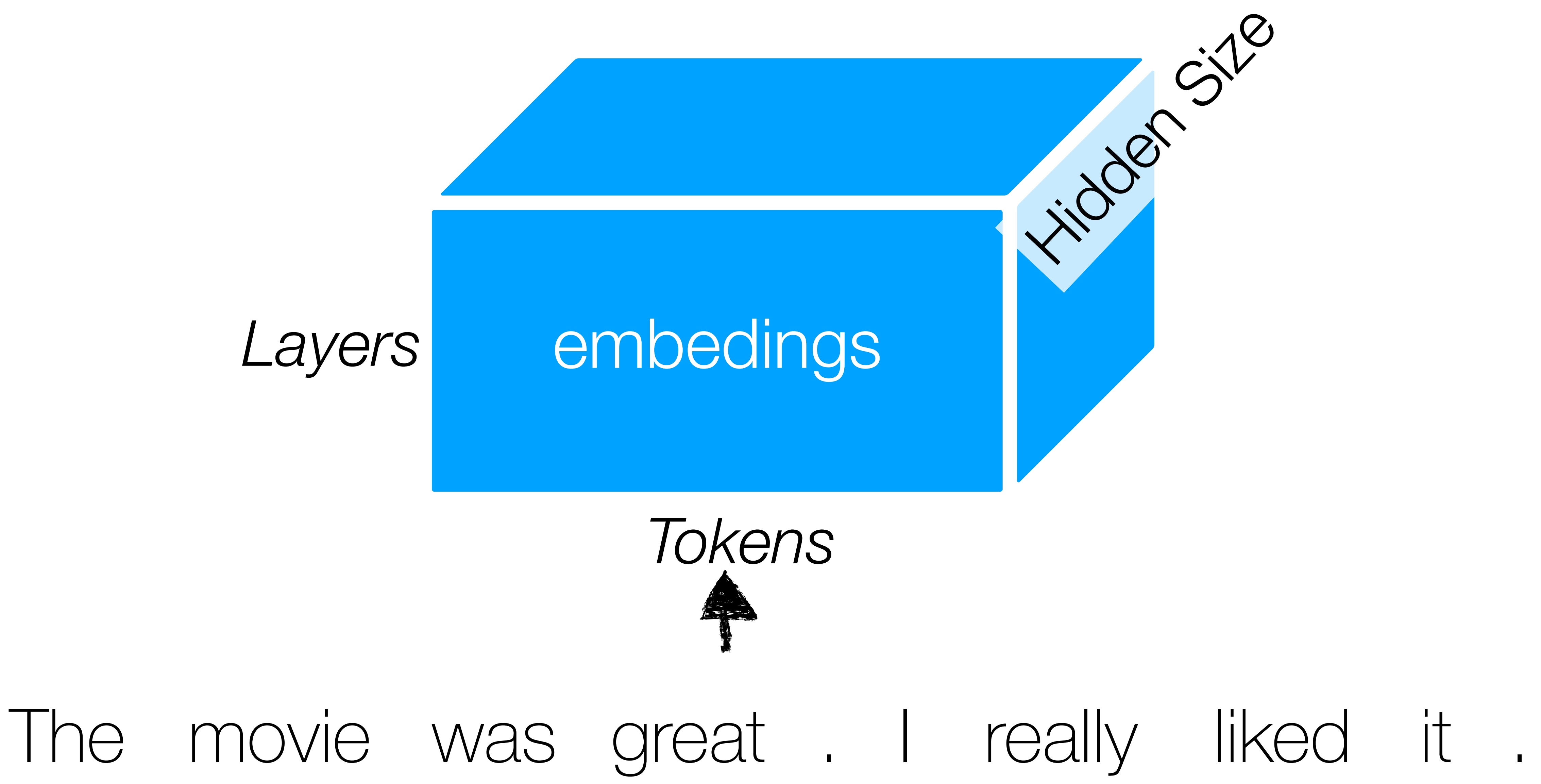
$$p = \min(\mathbb{P}(X \leq x),\ \mathbb{P}(X > x))$$
$$= \min(\mathbb{P}(X \leq x),\ 1 - \mathbb{P}(X \leq x))$$

x

MaSF

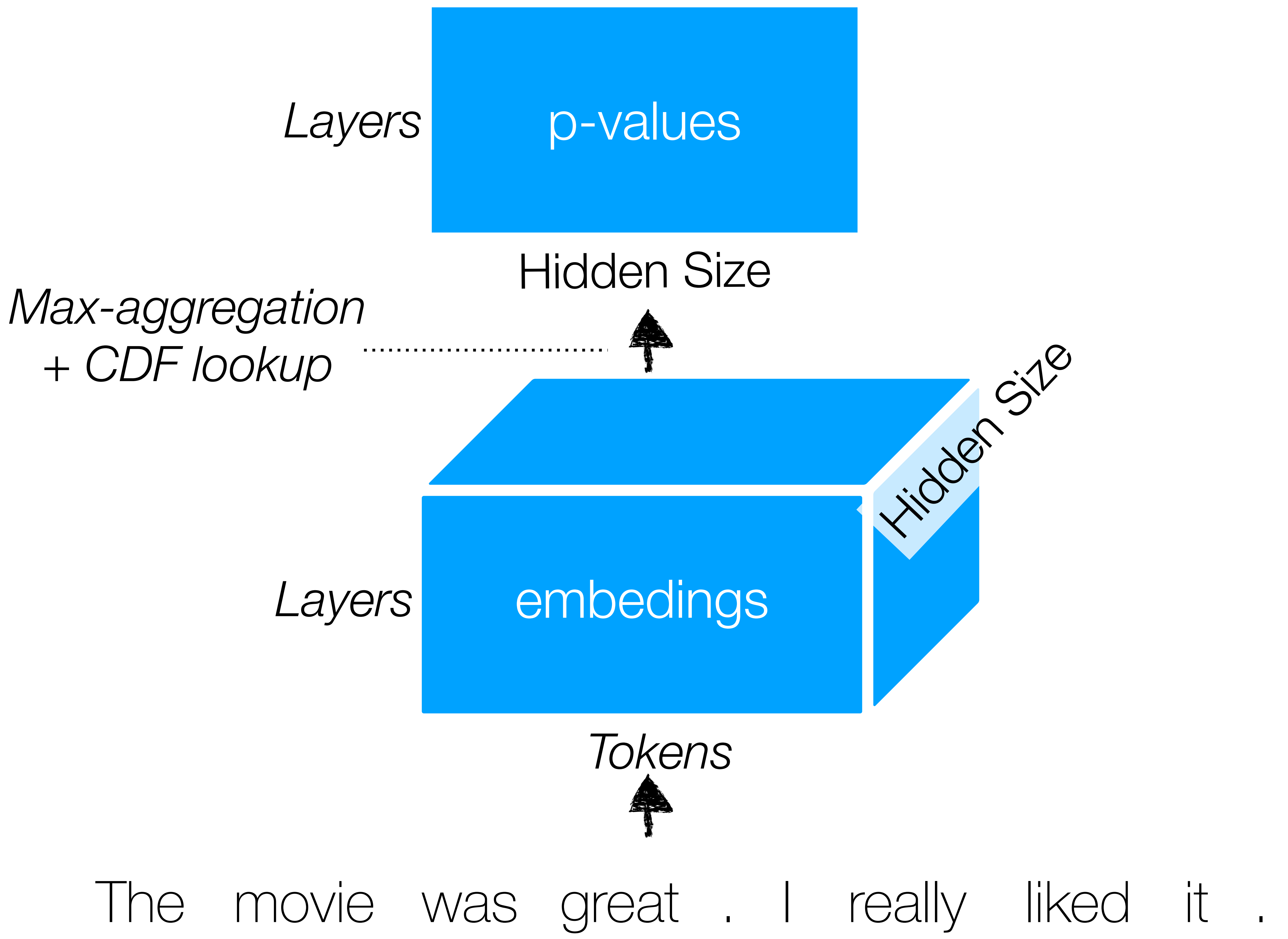
The movie was great . I really liked it .

MaSF

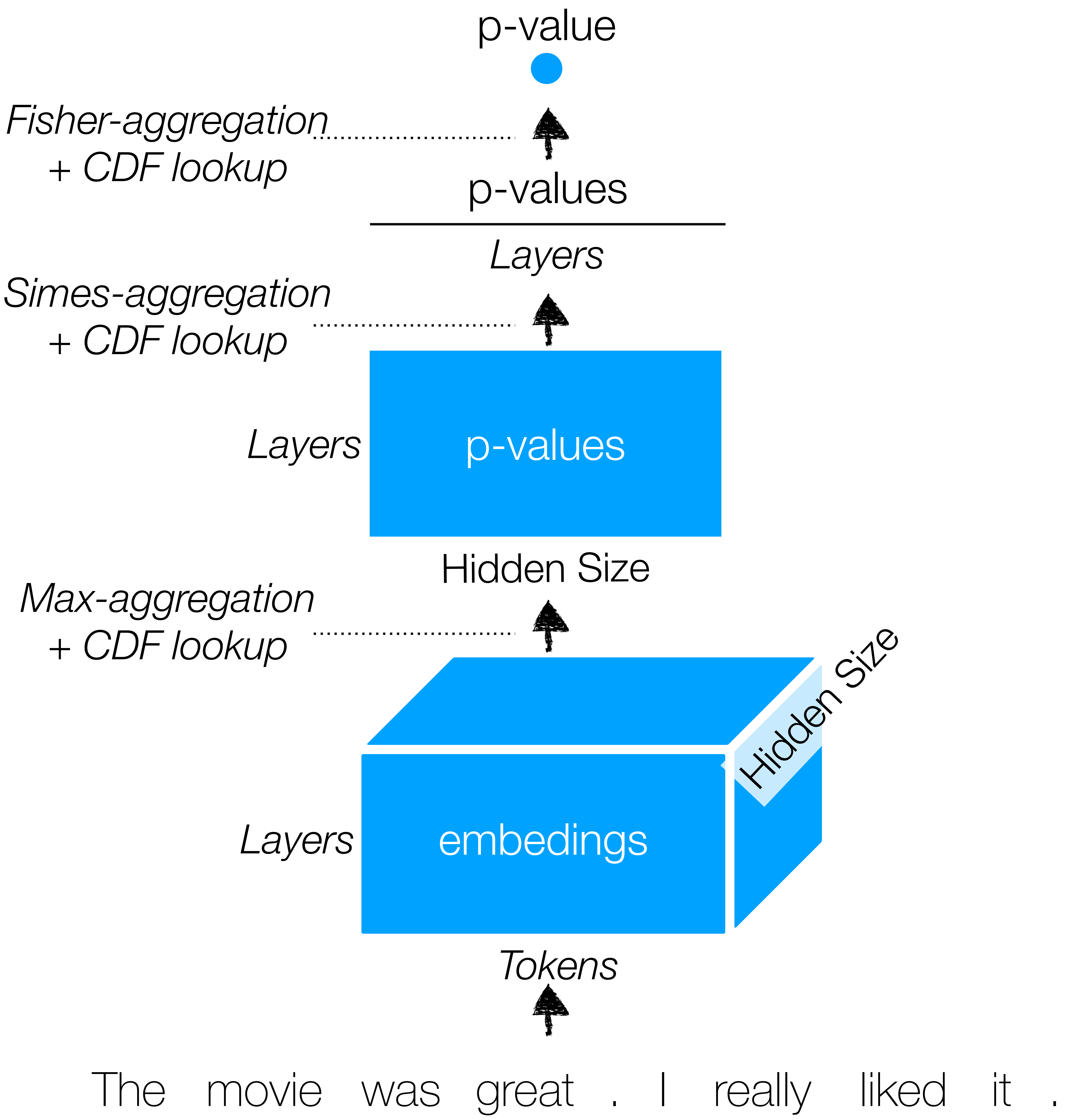


Hidden Size
Layers
embedings
Tokens
The movie was great . I really liked it .

MaSF


Layers
p-values
Hidden Size
Max-aggregation
+ CDF lookup
Hidden Size
Layers
embedings
Tokens
The movie was great . I really liked it .

MaSF


p-value
Fisher-aggregation
+ CDF lookup
p-values
Layers
Simes-aggregation
+ CDF lookup
Layers
p-values
Hidden Size
Max-aggregation
+ CDF lookup
Hidden Size
Layers
embedings
Tokens
The movie was great . I really liked it .

P-value aggregation

Bonferroni

Avoid p-fishing by dividing the threshold by N.

$$p_i < \frac{5\,\%}{N}$$

$$N \cdot \min_{i=1}^{N} p_i < 5\,\%$$

Simes

Consider all p-values.
For the smallest p-value (i=1) it is the same.

$$\min_{i=1}^{N} \frac{p_i \cdot N}{i} < 5\,\%$$

where $p_1 < p_2 < \cdots < p_N$

Fisher

No clear intuition. Follows a chi-squared distribution.

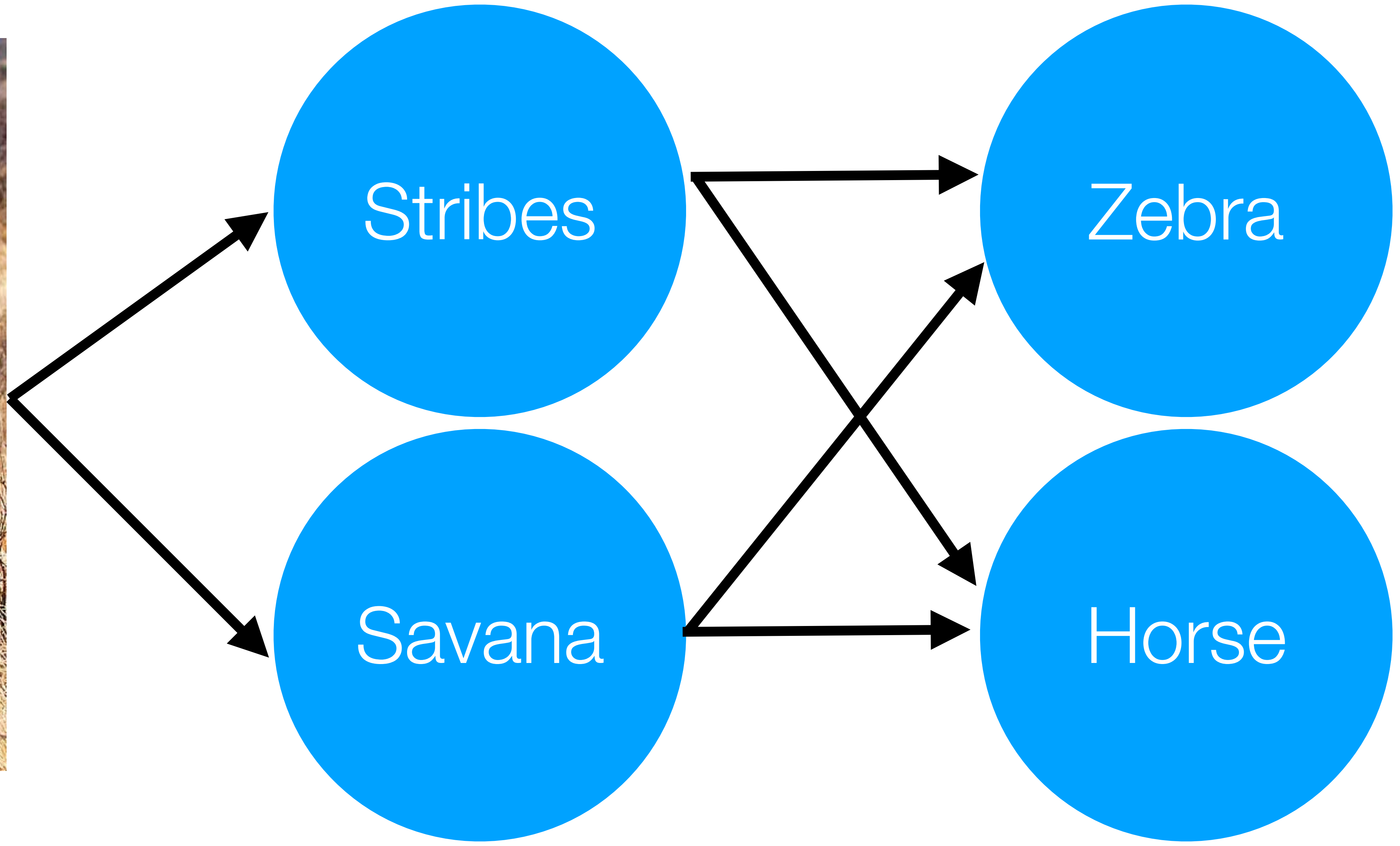
$$T = -2 \sum_{i=1}^{N} \ln(p_i)$$

FMMs for other explanations

Concept explanations

- Faithfulness of concepts is often measured using interventions in the intermediate state.
- These intervention likely cause out-of-distribution issues.

Grevy's Zebra Stallion, CC BY-SA 2.0



Self-explanations

Self-modeling

A model should be able to simulate itself, to explain itself in general.

Self-modeling

Meta-cognition question

Are you able to answer who was the first president of the United States? Yes/No

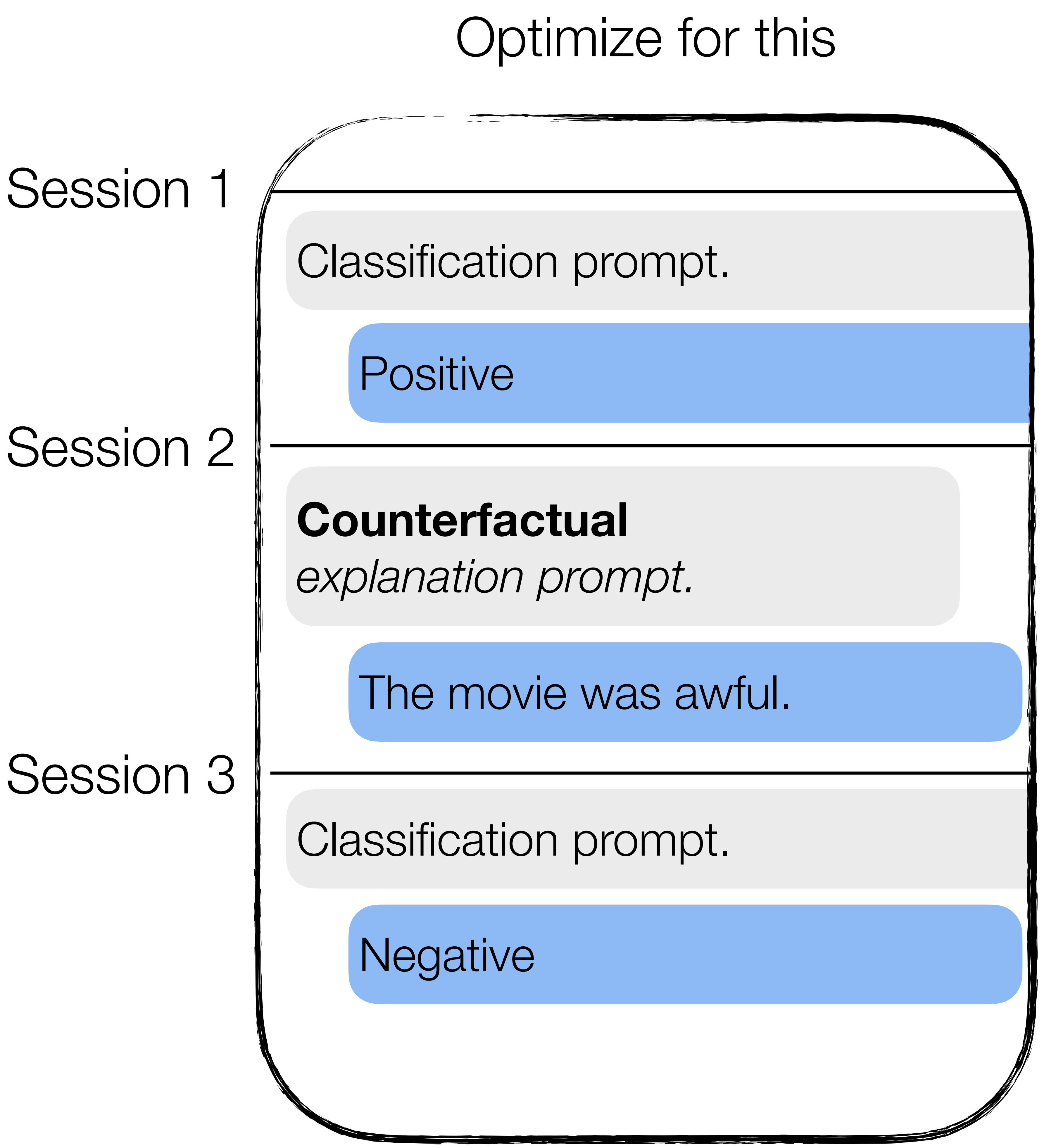
No

Direct question

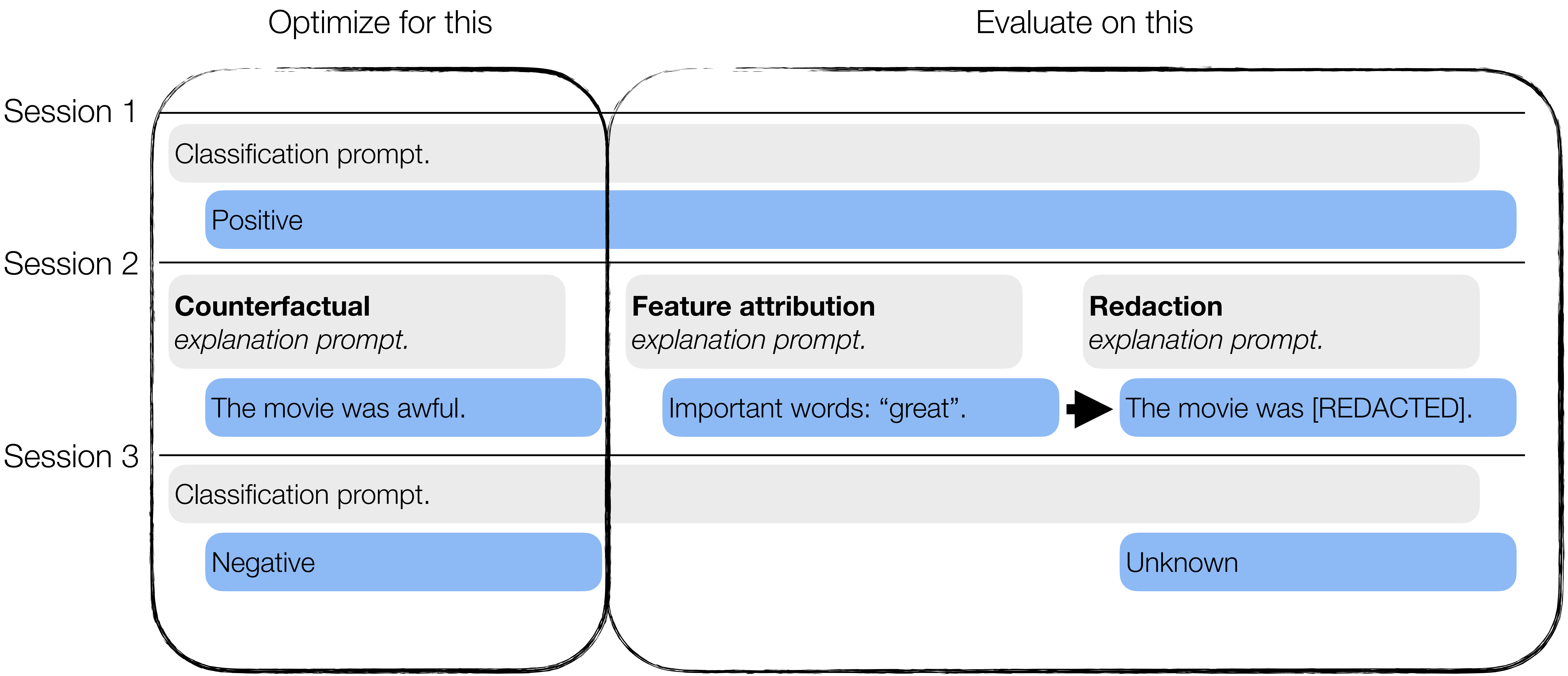
Who was the first president of the United States?

George Washington

How does this generalize?



How does this generalize?



On Measuring Faithfulness of Natural Language Explanations

Letitia Parcalabescu and **Anette Frank**
Computational Linguistics Department
Heidelberg University

Claims: currently no general faithfulness metric for natural language explanations

Abstract

Large language models (LLMs) can explain their own predictions, through post-hoc or Chain-of-Thought (CoT) explanations. However the LLM could make up reasonably sounding explanations that are unfaithful to its underlying reasoning. Recent work has designed tests that aim to judge the faithfulness of either post-h[illegible] CoT explanations. In this paper we ar[illegible] actually measuring faithfulness in terms of the models' inner workings, but only evaluate their self-consistency [illegible] of our work are two-fold. i) We aim to clarify the status of existing faithfulness tests in terms of [illegible] self-[illegible] we underline by constructing a *Comparative Consistency Bank* for self-consistency tests that for the first time compares existing tests on a common suite of 11 open-source LLMs and 5 datasets – including ii) our own proposed *self-consistency measure CC-SHAP*. CC-SHAP is a new fine-grained measure (not test) of LLM self-consistency that compares a model's input contributions to answer prediction and

et al., 2023), they can be surprisingly insensitive to the correctness of labels in in-context learning (Min et al., 2022) and can produce correct predictions even with irrelevant or misleading prompts (Webson and Pavlick, 2022).

Especially in cases of unintuitive behaviour, explanations for their way of acting would be helpful. Even though LLMs can provide plausibly sounding explanations for their answers, recent work argues [illegible] (NLEs) are often unfaithful (Atanasova et al., 2023; Lan[illegible]). [illegible] *faithful* explanations [illegible] *the true reasoning process of the model* (Jacovi and Goldberg, 2020) is imp[illegible] behind an AI system's answer and is instrumental for creating trustworthy AI. Being able to measure the faithfulness of an explanation is most critical when a model provides an answer we are unable to judge – whether it is AI uncovering new scientific facts or ChatGPT helping with homework.

Recent work aims to assess the faithfulness of LLM-produced NLEs through faithfulness tests (Atanasova et al., 2023; Turpin et al., 2023; Lan-

Integrated Gradient

Integrated Gradient axioms

Completeness

Attributions $\phi_i(x, f)$ for each feature i should sum to the total value $f(x)$.

$$\sum_{i=1}^{n} \phi_i(x, f) = f(x)$$

Implementation Invariance

The attributions are always identical for two functionally equivalent networks.

Sensitivity

If for every input and baseline that differ in one feature but have different predictions, then the differing feature should have non-zero attribution.

Integrated Gradient axioms

$$\mathbf{E}_{\text{integrated-gradient}}(\mathbf{x}, c) = (\mathbf{x} - \mathbf{b}) \odot \frac{1}{k} \sum_{i=1}^{k} \nabla_{\tilde{\mathbf{x}}_i} f(\tilde{\mathbf{x}}_i; \theta)_c, \quad \tilde{\mathbf{x}}_i = \mathbf{b} + {}^{i}\!/\!{}_{k}(\mathbf{x} - \mathbf{b}),$$

where $f(\mathbf{x}; \theta)$ is the model logits.

Shapely

Shapely axioms

Efficiency / Completeness

Attributions $\phi_i(x, f)$ for each player i should sum to the total value $f(x)$.

$$\sum_{i=1}^{n} \phi_i(x, f) = f(x)$$

Additivity / Linearity

If the value can be linearly decomposed a $f + g$, the attributions $\phi_i(x, f)$ can be decomposed too.

$$\phi_i(x, f + g) = \phi_i(x, f) + \phi_i(x, g)$$

Symmetry

If two players a and b are identical, they should receive equal attribution.

$$\phi_a(x, f) = \phi_b(x, f)$$
$$\text{if } f(S \cup \{a\}) = f(S \cup \{b\})$$
$$\forall S \subseteq x \backslash \{a, b\}$$

Null Player

Attribution for a player i who doesn't contribute is zero.

$$\phi_i(x, f) = 0$$
$$\text{if } f(S \cup \{i\}) = f(S)$$
$$\forall S \subseteq x \backslash \{i\}$$

Shapely

$$\phi_i(x, f) = \sum_{S \subseteq x \setminus \{i\}} \frac{|S|!\big(|x| - |S| - 1\big)!}{|x|!} \big(f(S \cup \{i\}) - f(S)\big)$$

Shapely Example

- $15 for Alice alone.
- Alice and Bob live together, but Bob wants a luxurious tax, adding 10$.
- Charlie lives further away, increases the cost to $51.

Passengers	Cost	Note
{∅}	$0	No taxi ride, no costs
{Alice}	$15	Standard fare to Alice’s & Bob’s place
{Bob}	$25	Bob always insists on luxury taxis
{Charlie}	$38	Charlie lives slightly further away
{Alice, Bob}	$25	Bob always gets his way
{Alice, Charlie}	$41	Drop off Alice first, then Charlie
{Bob, Charlie}	$51	Drop off luxurious Bob first, then Charlie
{Alice, Bob, Charlie}	$51	The full fare with all three of them

Shapely Example

1. Consider every order of Alice, Bob, Charlie.

- Alice, Bob, Charlie
- Alice, Charlie, Bob
- Bob, Alice, Charlie
- Charlie, Alice, Bob
- Bob, Charlie, Alice
- Charlie, Bob, Alice

Passengers	Cost
{∅}	$0
{Alice}	$15
{Bob}	$25
{Charlie}	$38
{Alice, Bob}	$25
{Alice, Charlie}	$41
{Bob, Charlie}	$51
{Alice, Bob, Charlie}	$51

Shapely Example

1. Consider every order of Alice, Bob, Charlie.

2. Consider Alice is the last to enter the taxi.

Passengers	Cost
{∅}	$0
{Alice}	$15
{Bob}	$25
{Charlie}	$38
{Alice, Bob}	$25
{Alice, Charlie}	$41
{Bob, Charlie}	$51
{Alice, Bob, Charlie}	$51

- Alice, Bob, Charlie
- Alice, Charlie, Bob
- Bob, Alice, Charlie
- Charlie, Alice, Bob
- Bob, Charlie, Alice
- Charlie, Bob, Alice

Shapely Example

1. Consider every order of Alice, Bob, Charlie.

2. Consider Alice is the last to enter the taxi.

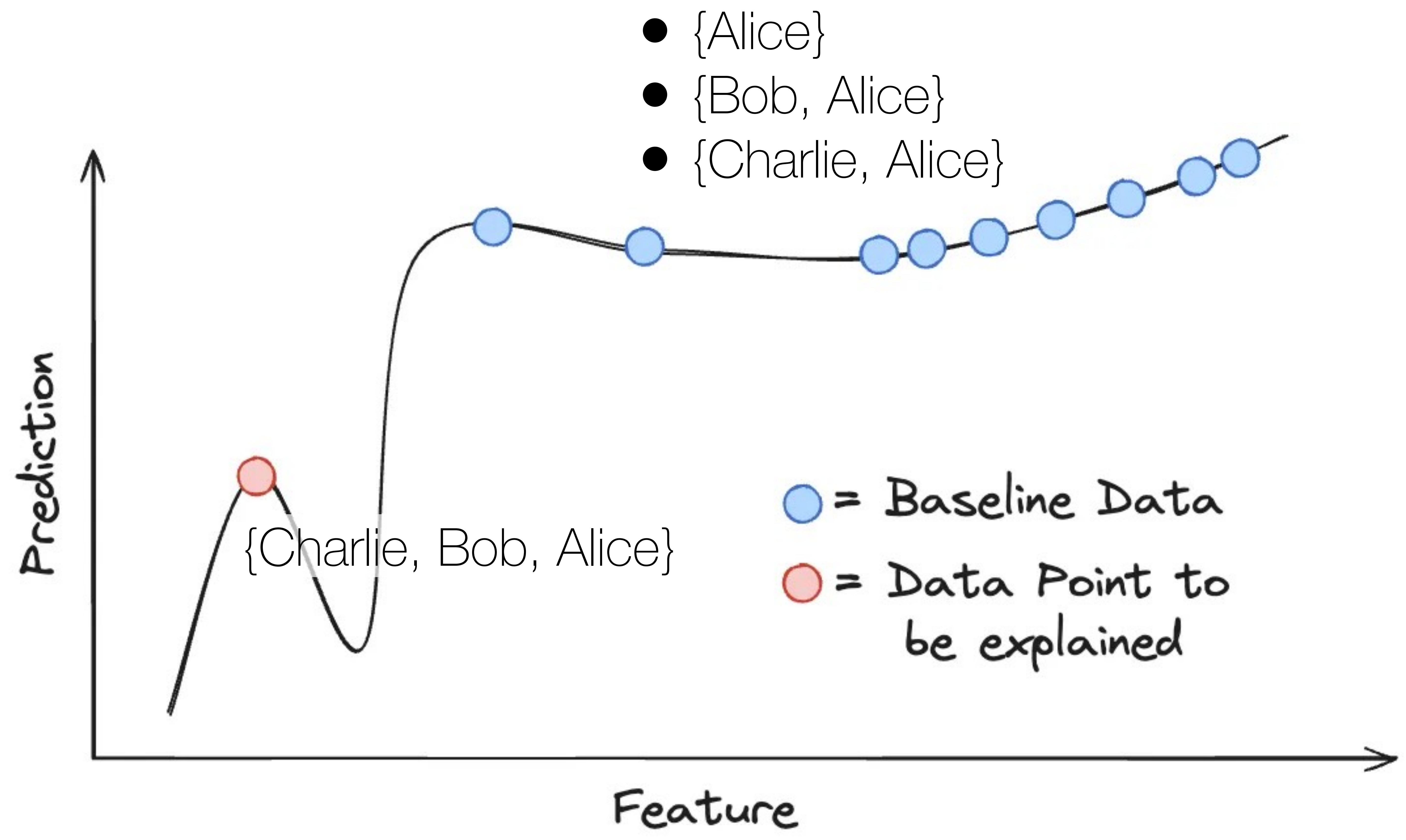
3. Average up Alice's contributions.

Passengers	Cost
{∅}	\$0
{Alice}	\$15
{Bob}	\$25
{Charlie}	\$38
{Alice, Bob}	\$25
{Alice, Charlie}	\$41
{Bob, Charlie}	\$51
{Alice, Bob, Charlie}	\$51

- Alice, Bob, Charlie — {∅} → {Alice} = \$15
- Alice, Charlie, Bob — {∅} → {Alice} = \$15
- Bob, Alice, Charlie — {Bob} → {Alice, Bob} = \$0
- Charlie, Alice, Bob — {Charlie} → {Alice, Charlie} = \$3
- Bob, Charlie, Alice — {Bob, Charlie} → {Alice, Bob, Charlie} = \$0
- Charlie, Bob, Alice — {Bob, Charlie} → {Alice, Bob, Charlie} = \$0

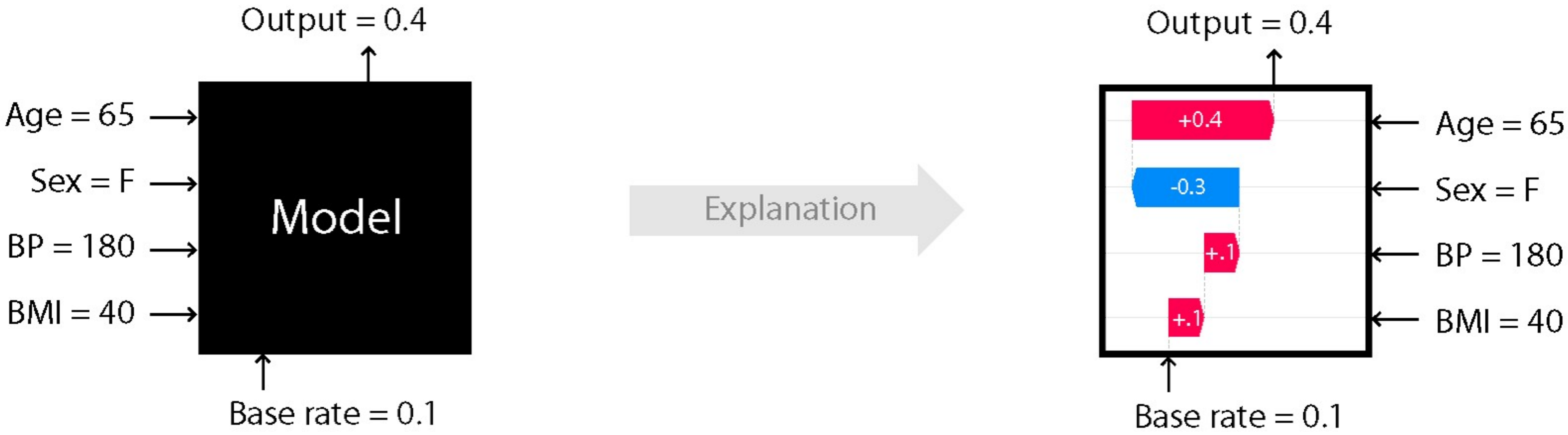
Average: \$5.5

Background / Baseline data



https://mindfulmodeler.substack.com/p/shedding-light-on-impossibility-theorems

Visualization



https://shap.readthedocs.io/en/latest/