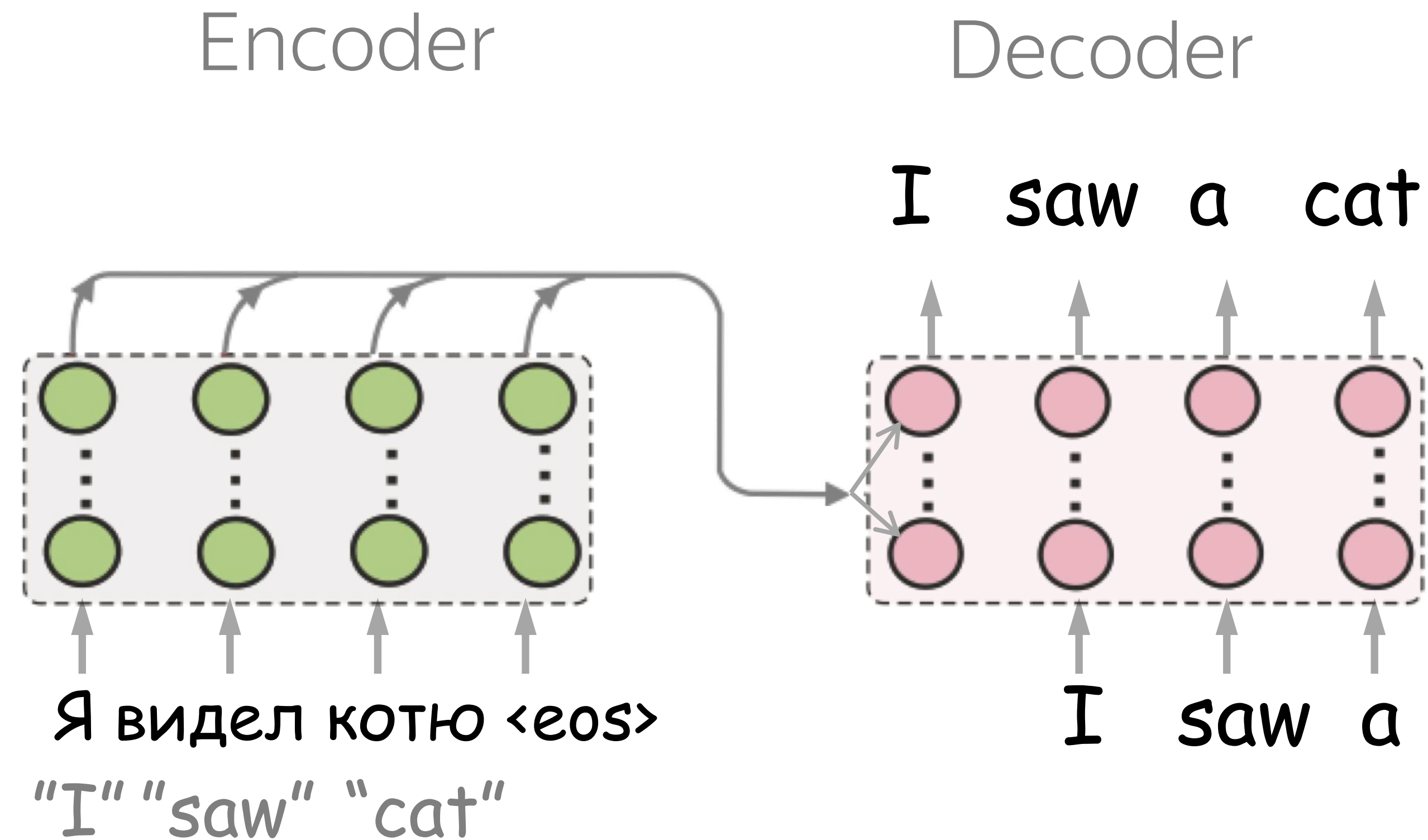


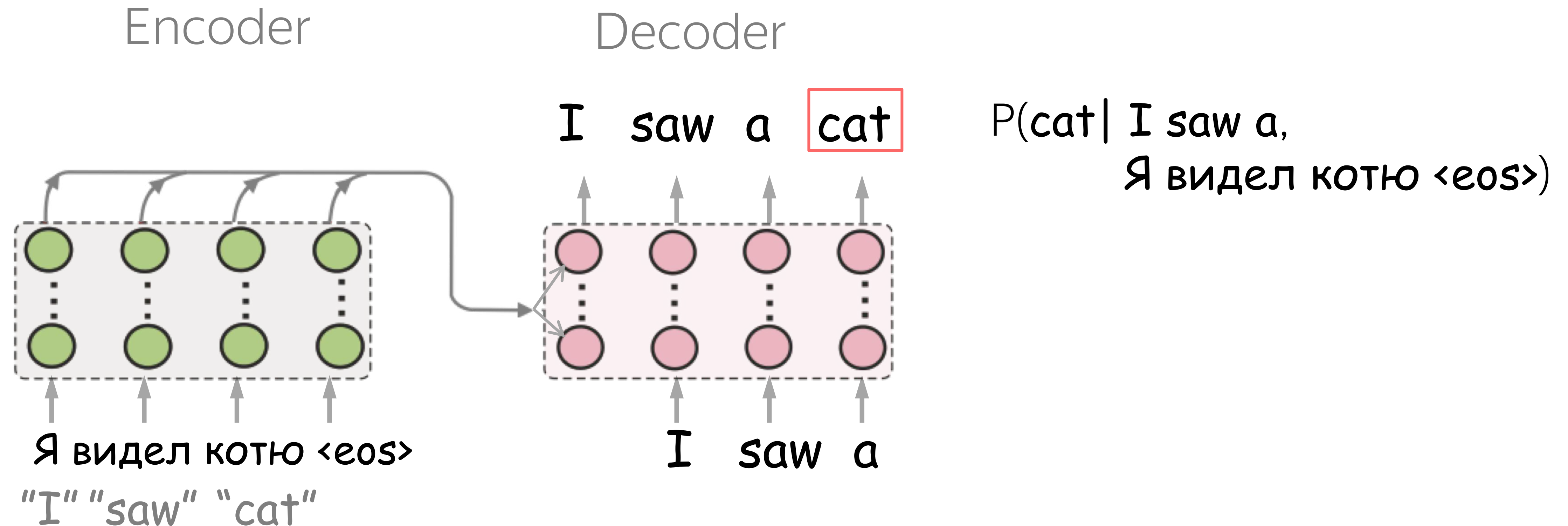
# NMT Analysis: The Trade-Off Between Source and Target, and (a bit of) the Training Process

Lena Voita<sup>1,2</sup>   Rico Sennrich<sup>3,1</sup>   Ivan Titov<sup>1,2</sup>

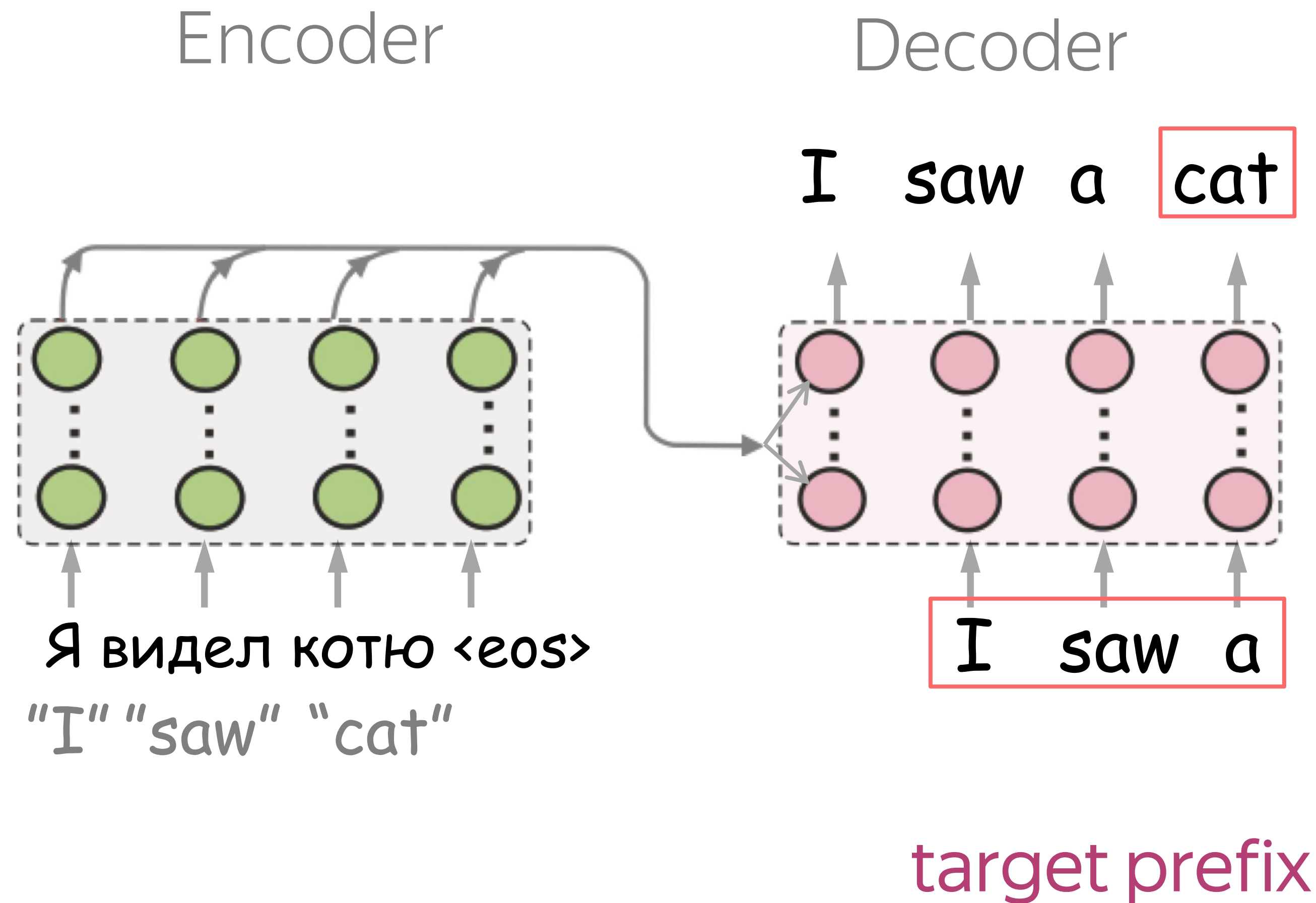
# NMT: The Trade-Off Between Source and Target



# NMT: The Trade-Off Between Source and Target



# NMT: The Trade-Off Between Source and Target

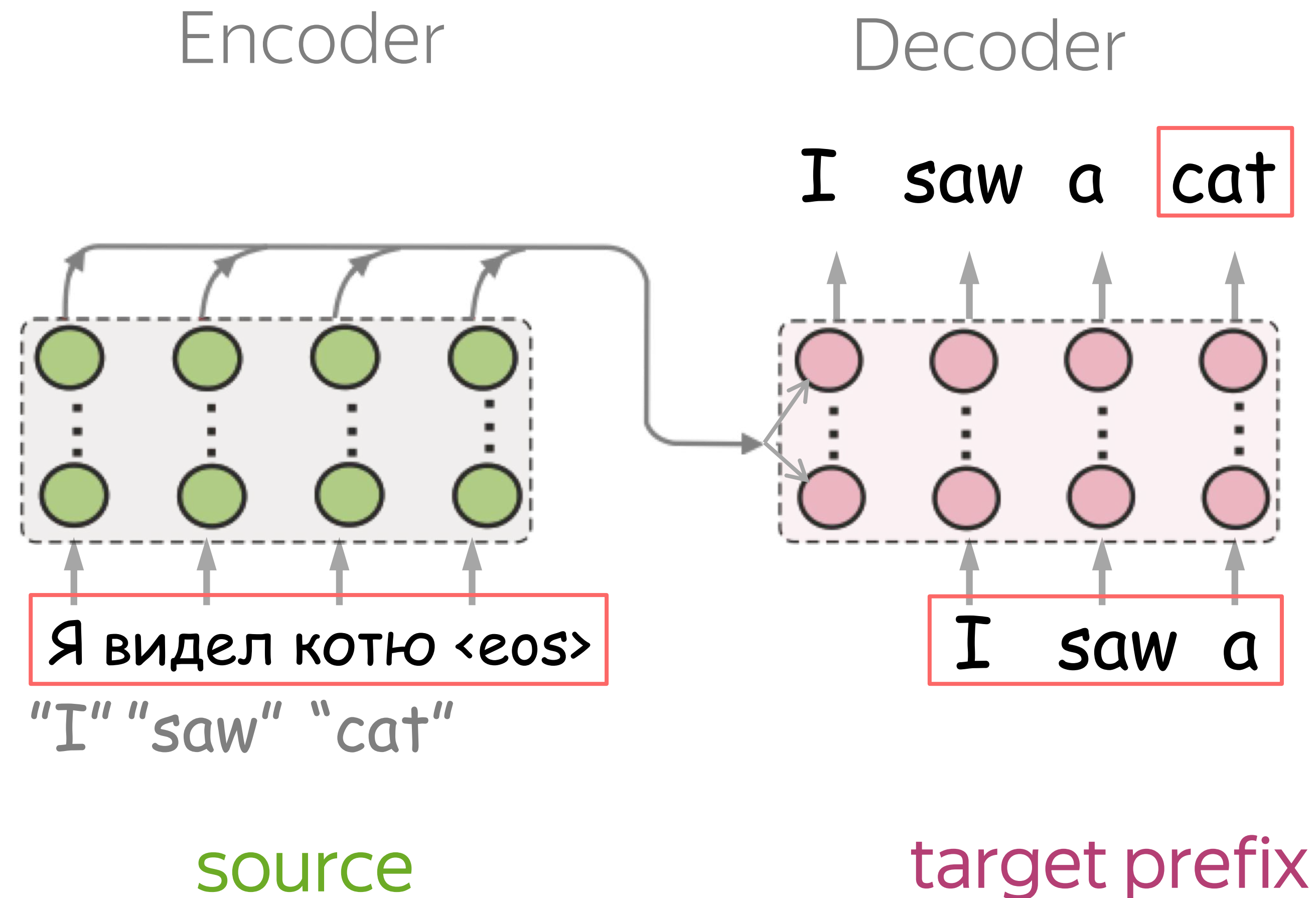


$P(\text{cat} | \text{I saw a}, \text{Я видел котю} \langle \text{eos} \rangle)$

Two types of context:

- target prefix

# NMT: The Trade-Off Between Source and Target



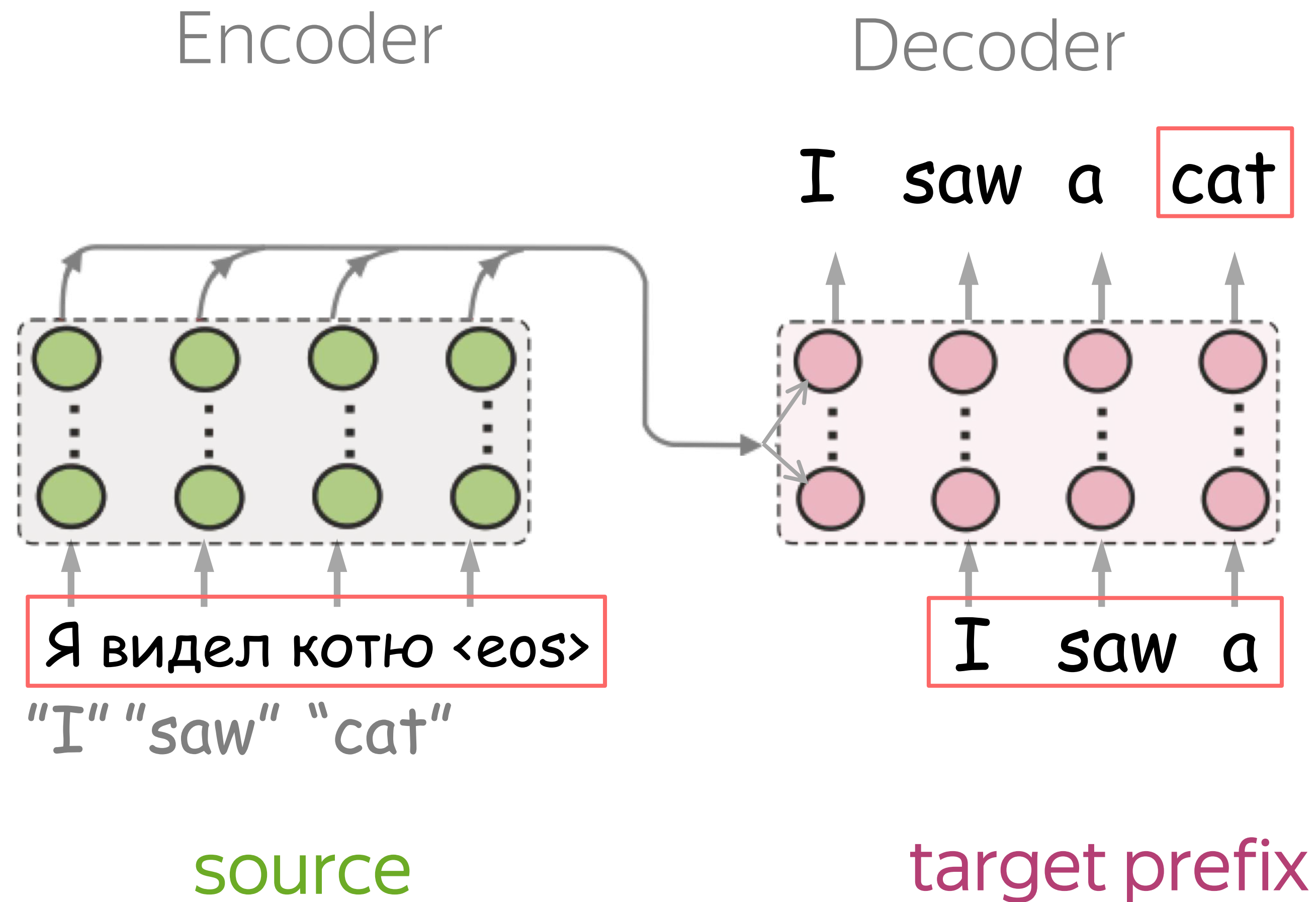
$P(\text{cat} | \text{I saw a, Я видел котю <eos>})$

Two types of context:

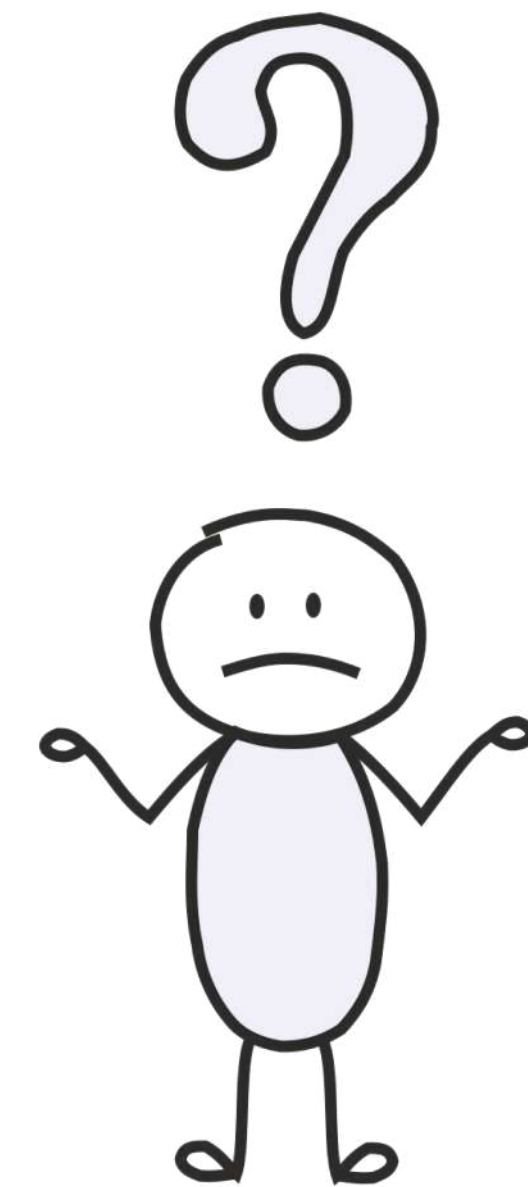
- target prefix
- source



# NMT: The Trade-Off Between Source and Target



What influences predictions:  
source or target?



# Why this is important: models fail to effectively use source and target

- context gates which weigh source and target contexts help in both RNNs (Tu et al., 2017; Wang et al., 2018) and Transformer (Li et al., 2020)

# Why this is important: models fail to effectively use source and target

- context gates which weigh source and target contexts help in both RNNs (Tu et al., 2017; Wang et al., 2018) and Transformer (Li et al., 2020)
- when hallucinating, a model fails to use source (Lee et al, 2018, Berard et al., 2019)

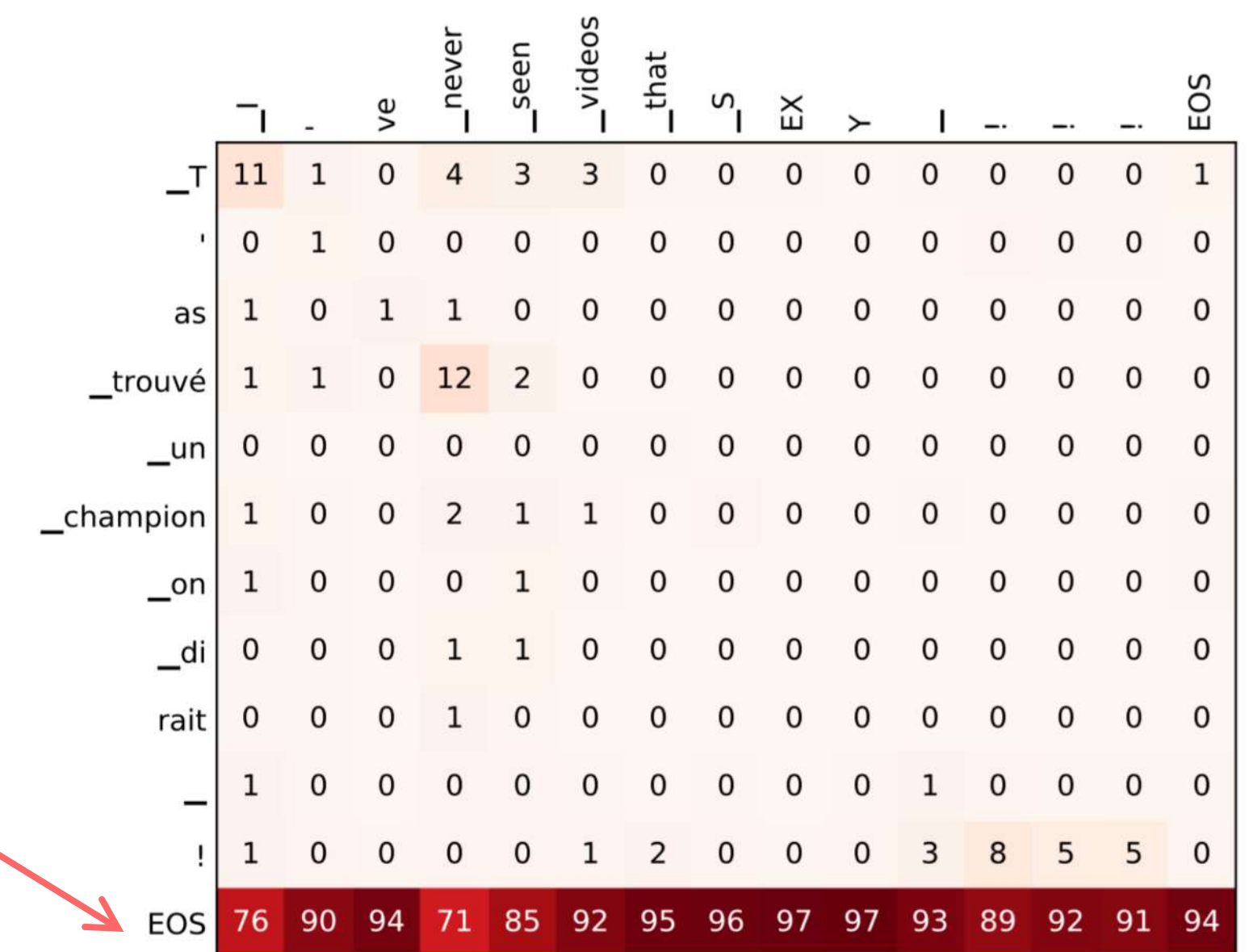


# Why this is important: models fail to effectively use source and target

- context gates which weigh source and target contexts help in both RNNs (Tu et al., 2017; Wang et al., 2018) and Transformer (Li et al., 2020)
- when hallucinating, a model fails to use source (Lee et al., 2018, Berard et al., 2019)

Evidence is based on heuristics.

E.g., most of attention is concentrated on source EOS



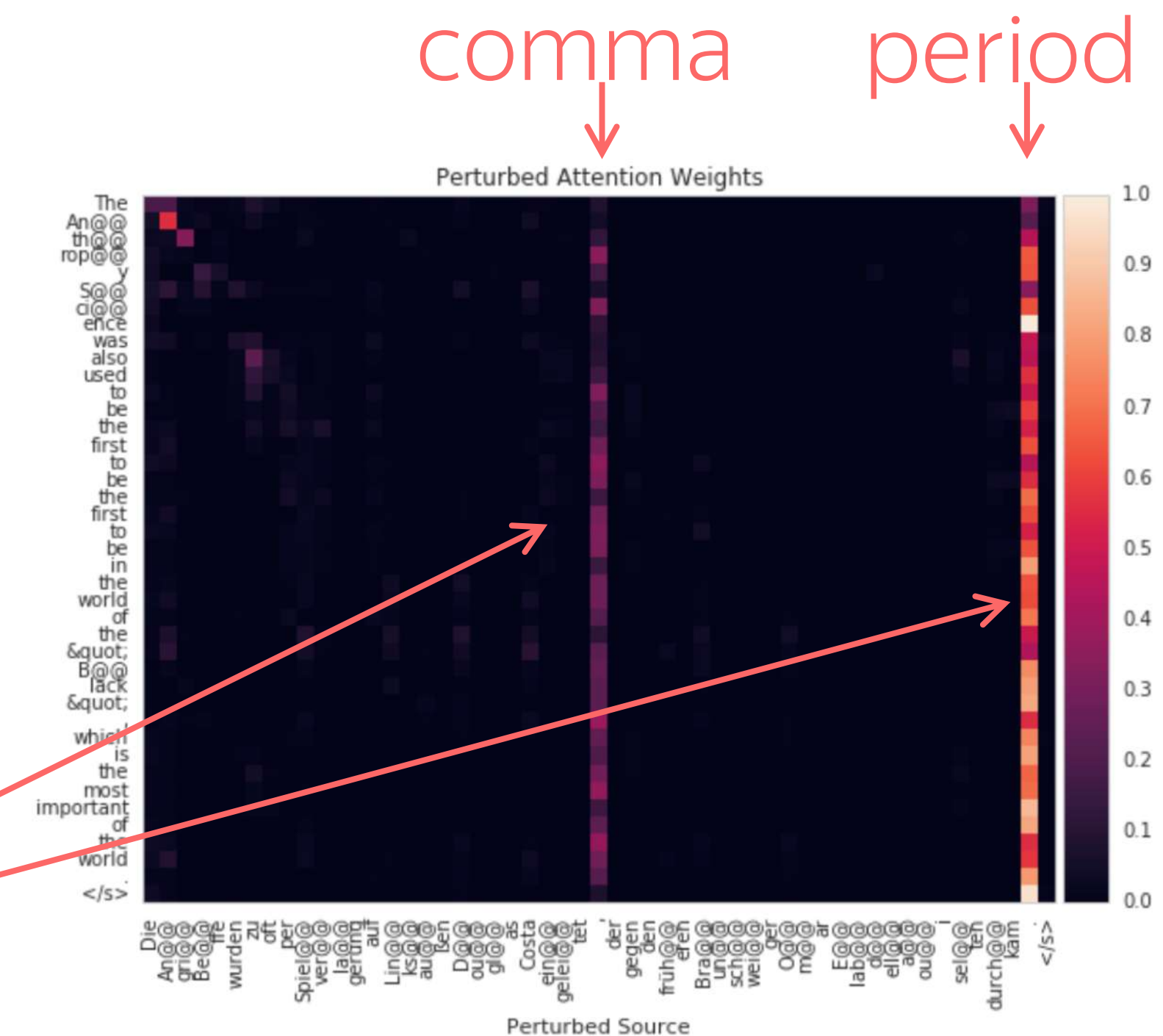
	_T	'	ve	_never	_seen	_videos	_that	_S	EX	Y	_	_	_	_	EOS
_T	11	1	0	4	3	3	0	0	0	0	0	0	0	0	1
'	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
as	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0
_trouv�	1	1	0	12	2	0	0	0	0	0	0	0	0	0	0
_un	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
_champion	1	0	0	2	1	1	0	0	0	0	0	0	0	0	0
_on	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
_di	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
rait	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
_	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
!	1	0	0	0	0	1	2	0	0	0	3	8	5	5	0
EOS	76	90	94	71	85	92	95	96	97	97	93	89	92	91	94

# Why this is important: models fail to effectively use source and target

- context gates which weigh source and target contexts help in both RNNs (Tu et al., 2017; Wang et al., 2018) and Transformer (Li et al., 2020)
- when hallucinating, a model fails to use source (Lee et al., 2018, Berard et al., 2019)

Evidence is based on heuristics.

E.g., most of attention is concentrated on source EOS or only on a few frequent tokens



Picture from: Lee et al., 2018

# This can also be useful for other applications

A method which estimates how a model uses source may be useful to

- evaluate techniques which force a model to rely on input (e.g., regularizations, additional loss terms, etc.)
- evaluate models for other tasks where reliance on source is important (e.g., data to text generation, image captioning, etc.)

# What is going to happen:

## The Trade-Off Between Source and Target

- Our Approach: (a version of) LRP
- Experiments
  - Getting Acquainted
  - Reference, Model and Random Prefixes
  - Exposure Bias and Source Contribution
  - Varying the Amount of Data
  - Training Stages

(A Bit of) the Training Process (work in progress)

# What is going to happen:

## The Trade-Off Between Source and Target

- Our Approach: (a version of) LRP
- Experiments
  - Getting Acquainted
  - Reference, Model and Random Prefixes
  - Exposure Bias and Source Contribution
  - Varying the Amount of Data
  - Training Stages

(A Bit of) the Training Process (work in progress)

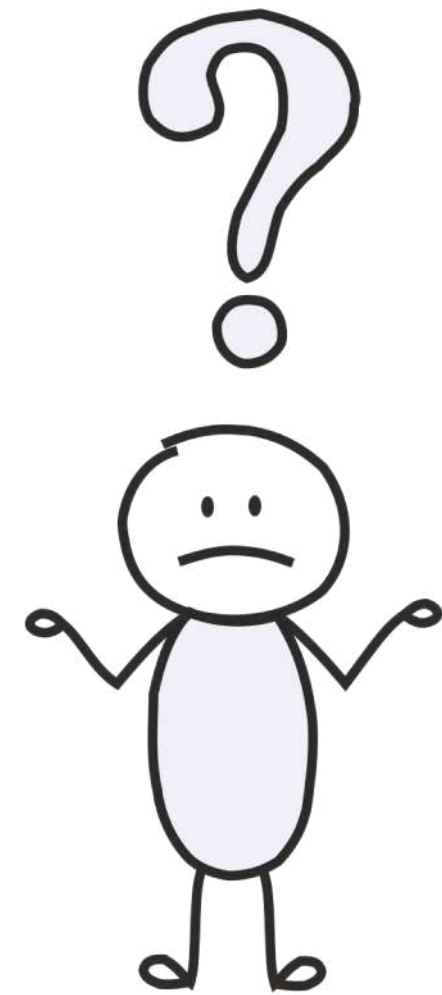


# Our Approach:

## Layerwise Relevance Propagation

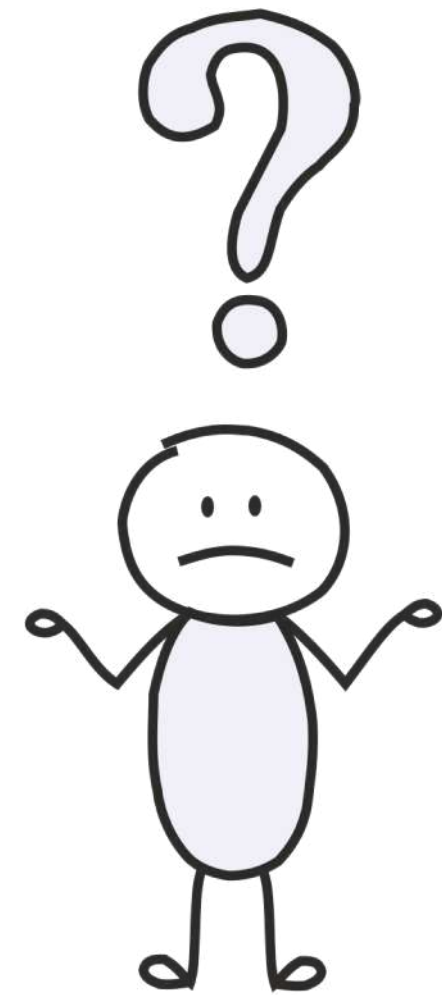


# What do we want?



What influences  
predictions:  
source or target?

# What do we want?

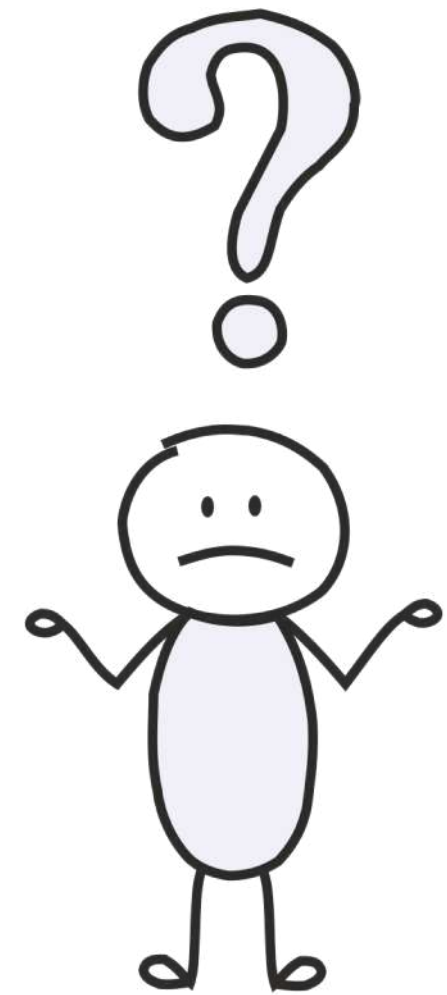


What influences  
predictions:  
**source** or **target**?

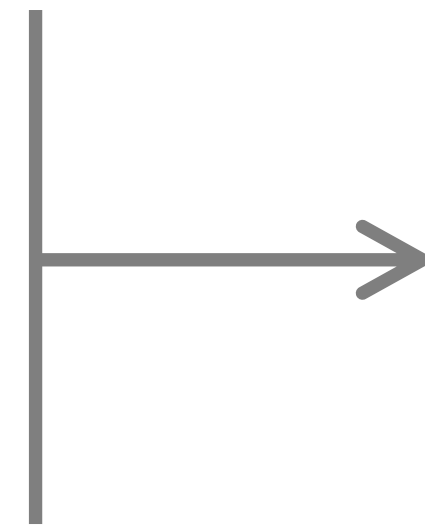
Previous work:

- tried to evaluate which individual tokens influence a prediction  
(e.g., Alvarez-Melis & Jaakkola, EMNLP 2017, He et al, EMNLP 2019)
- these influences were abstract quantities and did not reflect part of the total contribution

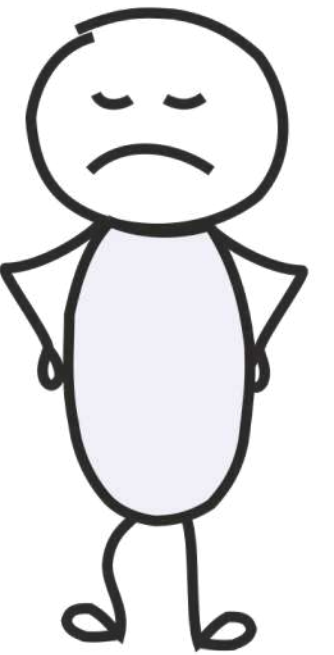
# What do we want?



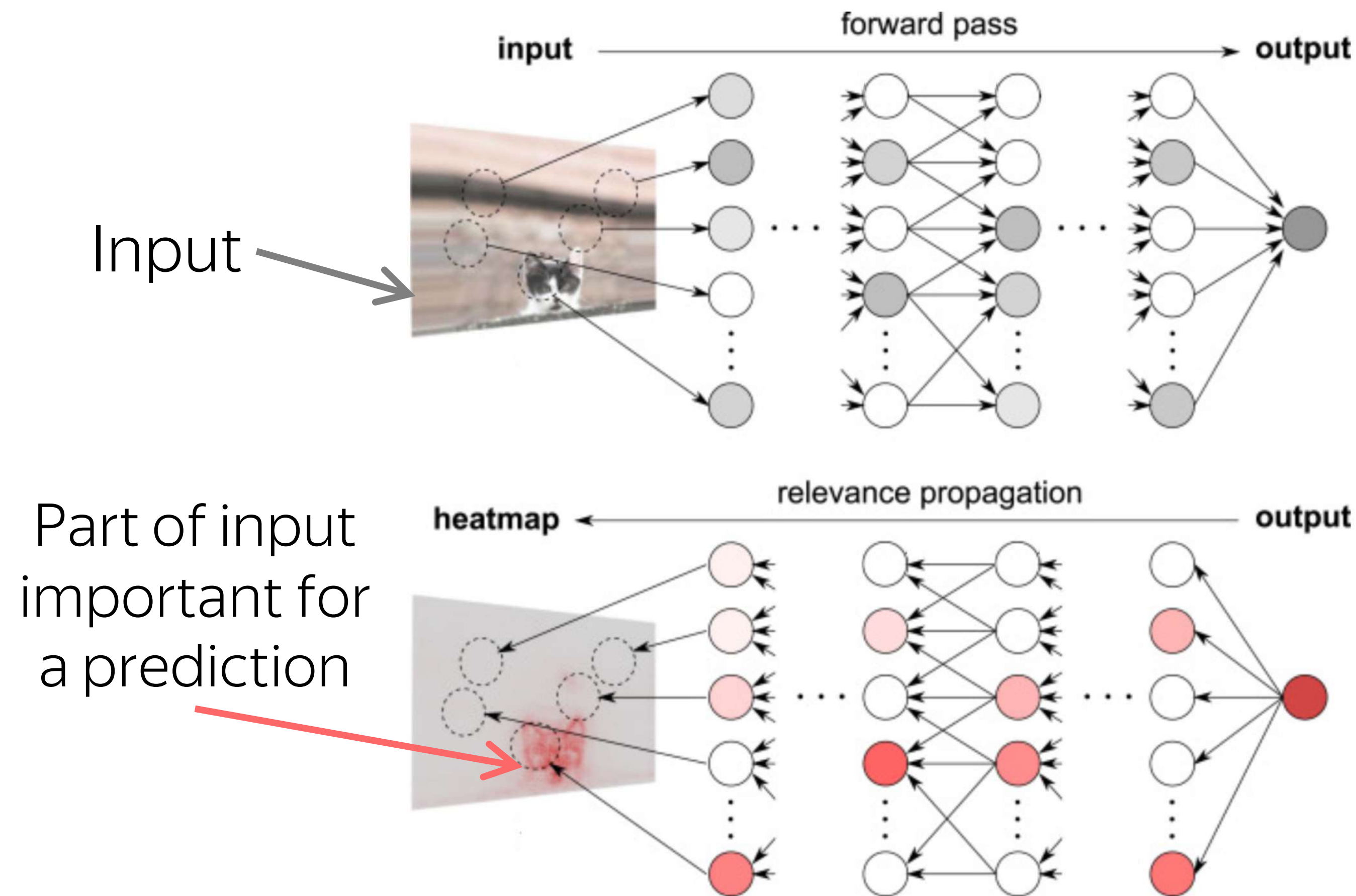
What influences  
predictions:  
**source** or **target**?



We want: not an  
abstract quantity, but  
**relative** contributions  
(i.e., part of the  
total contribution)



# Layerwise Relevance Propagation

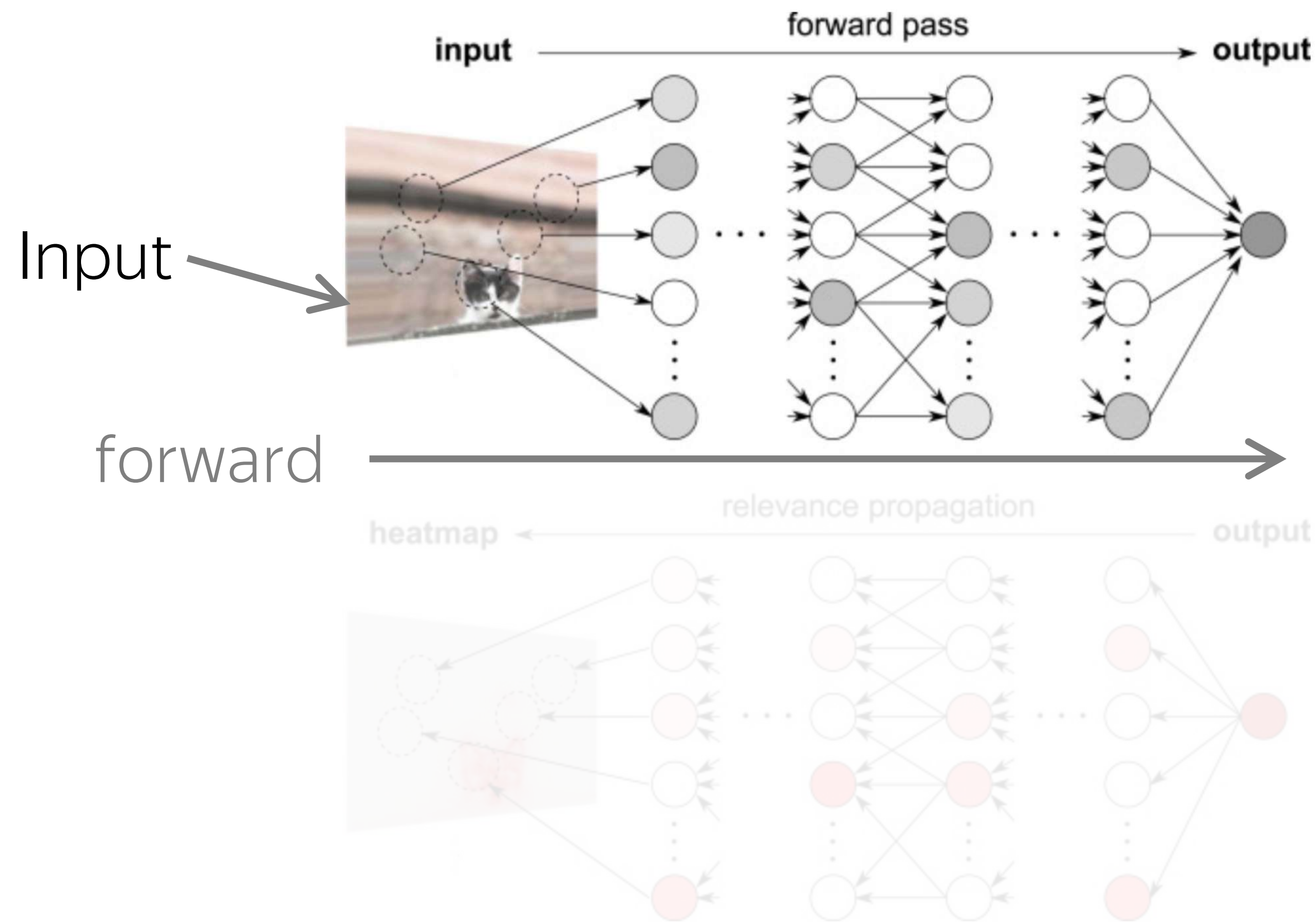


- identifies which pixels contributed to a prediction
- back-propagates relevance recursively from the output layer to the input

Illustration from: <http://danshiebler.com/2017-04-16-deep-taylor-lrp/>



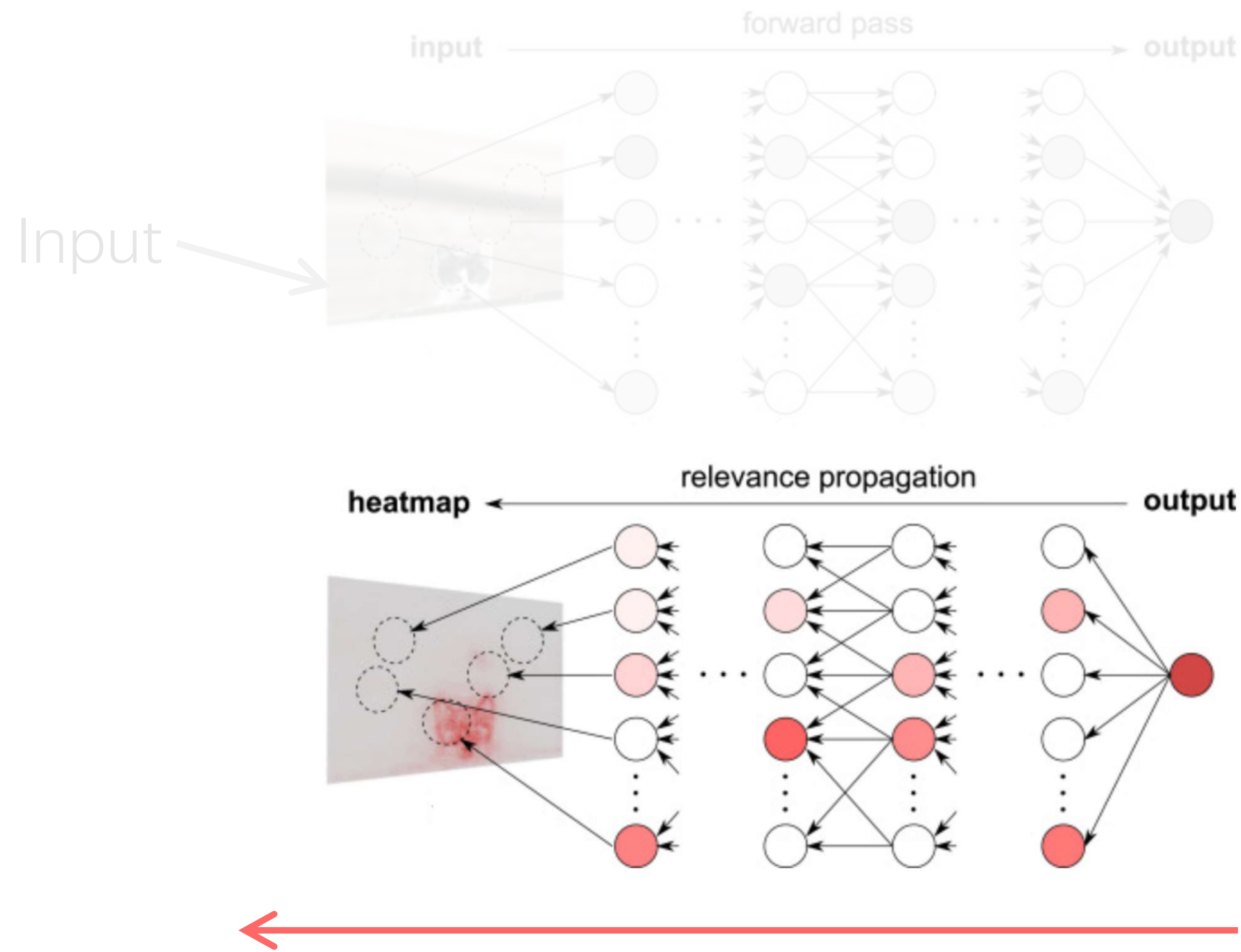
# Layerwise Relevance Propagation



- identifies which pixels contributed to a prediction
- back-propagates relevance recursively from the output layer to the input

Illustration from: <http://danshiebler.com/2017-04-16-deep-taylor-lrp/>

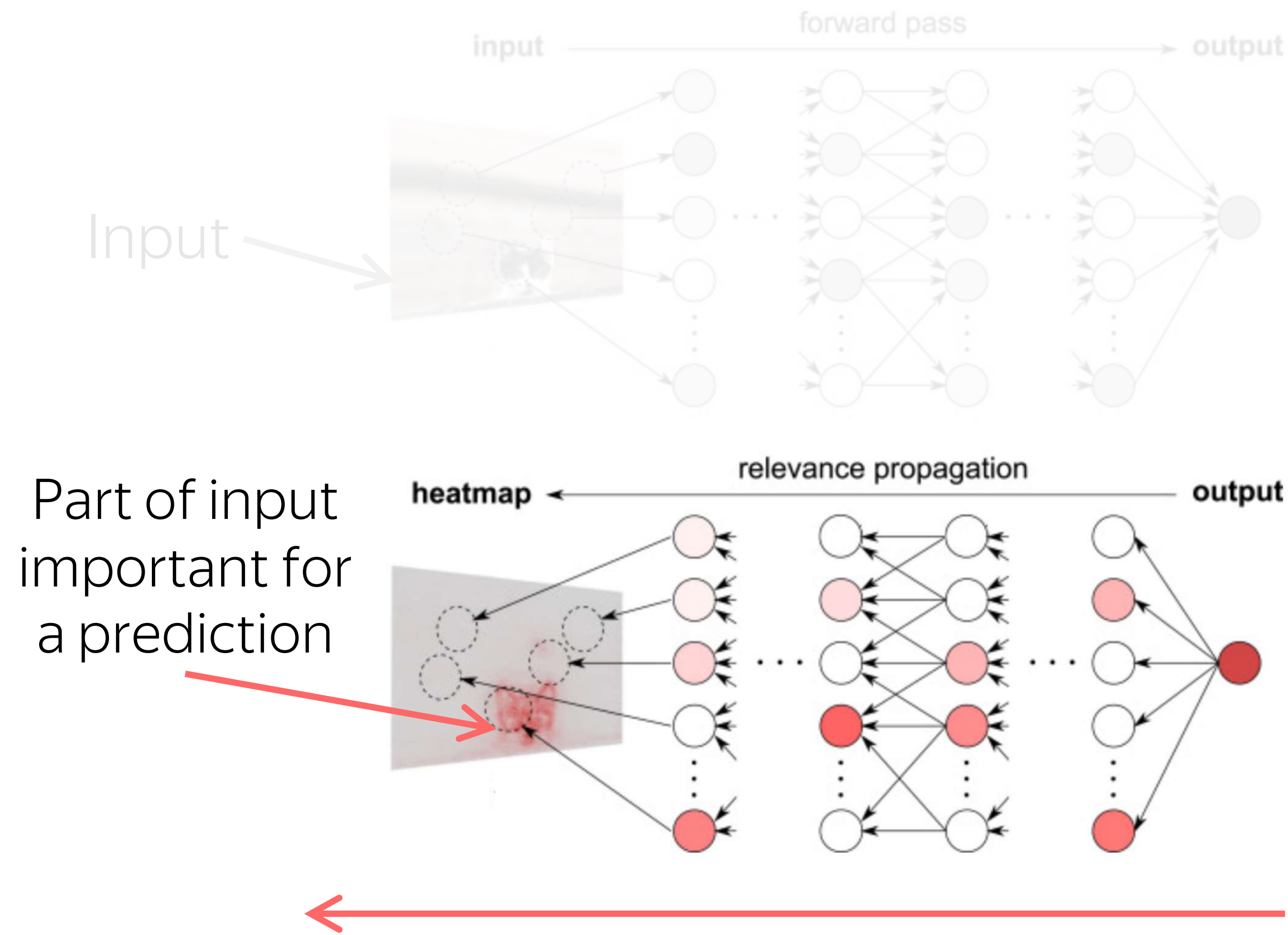
# Layerwise Relevance Propagation



- identifies which pixels contributed to a prediction
- back-propagates relevance recursively from the output layer to the input

Illustration from: <http://danshiebler.com/2017-04-16-deep-taylor-lrp/>

# Layerwise Relevance Propagation

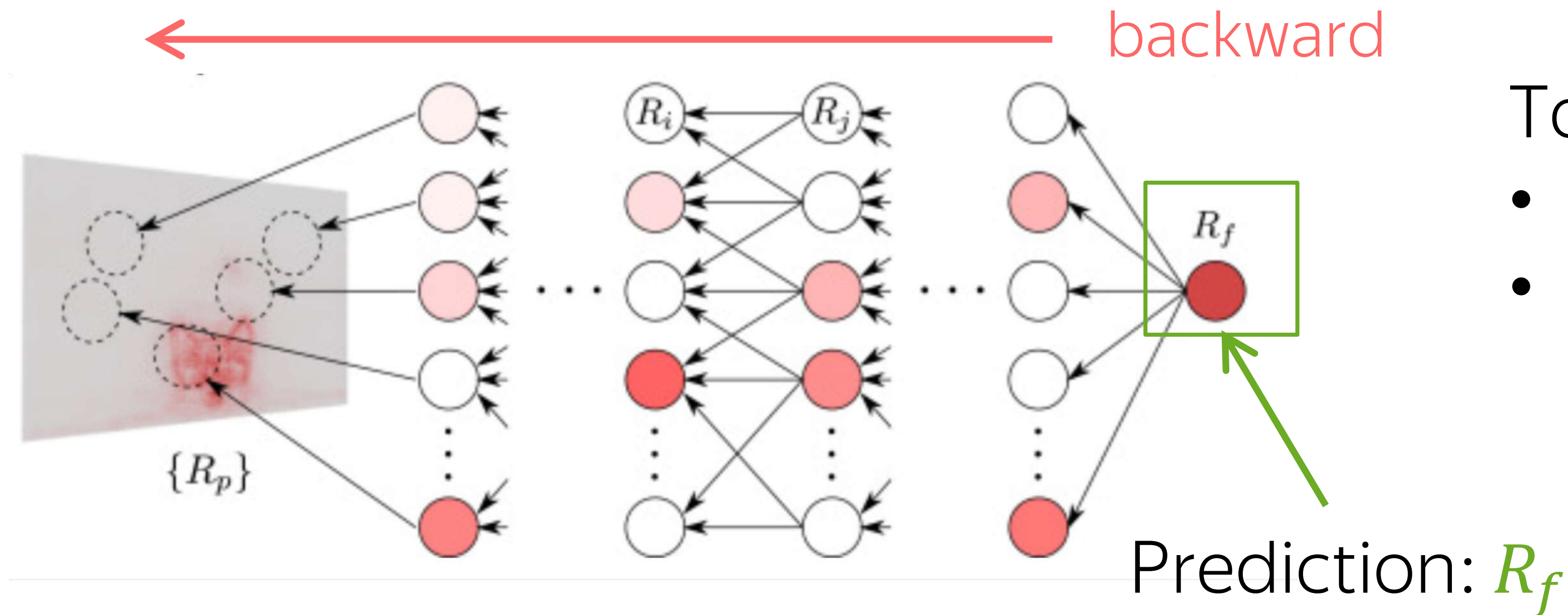


- identifies which pixels contributed to a prediction
- back-propagates relevance recursively from the output layer to the input

Illustration from: <http://danshiebler.com/2017-04-16-deep-taylor-lrp/>

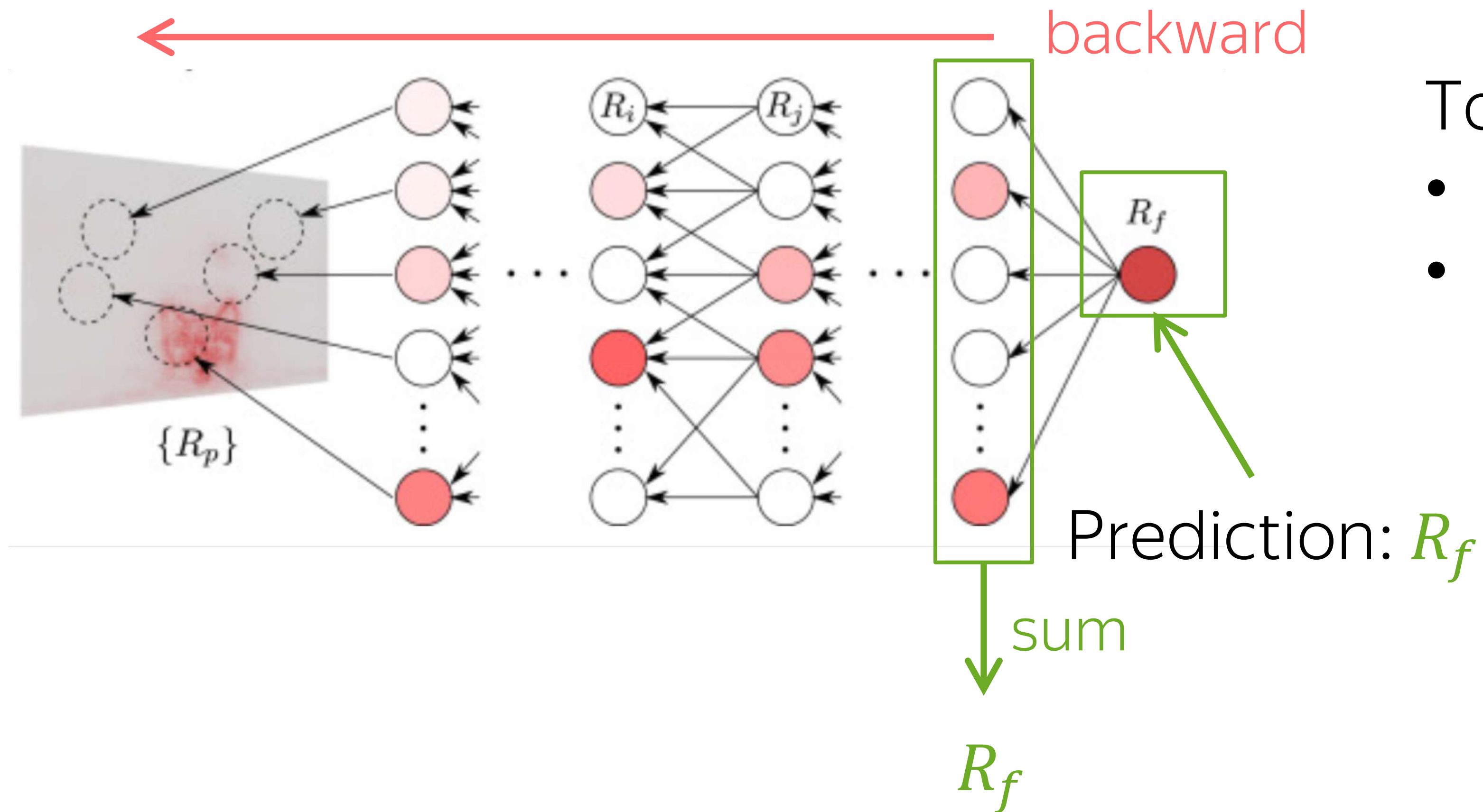


# Why LRP: Layerwise Conservation Principle



- Total relevance is
- constant
  - equals the prediction

# Why LRP: Layerwise Conservation Principle



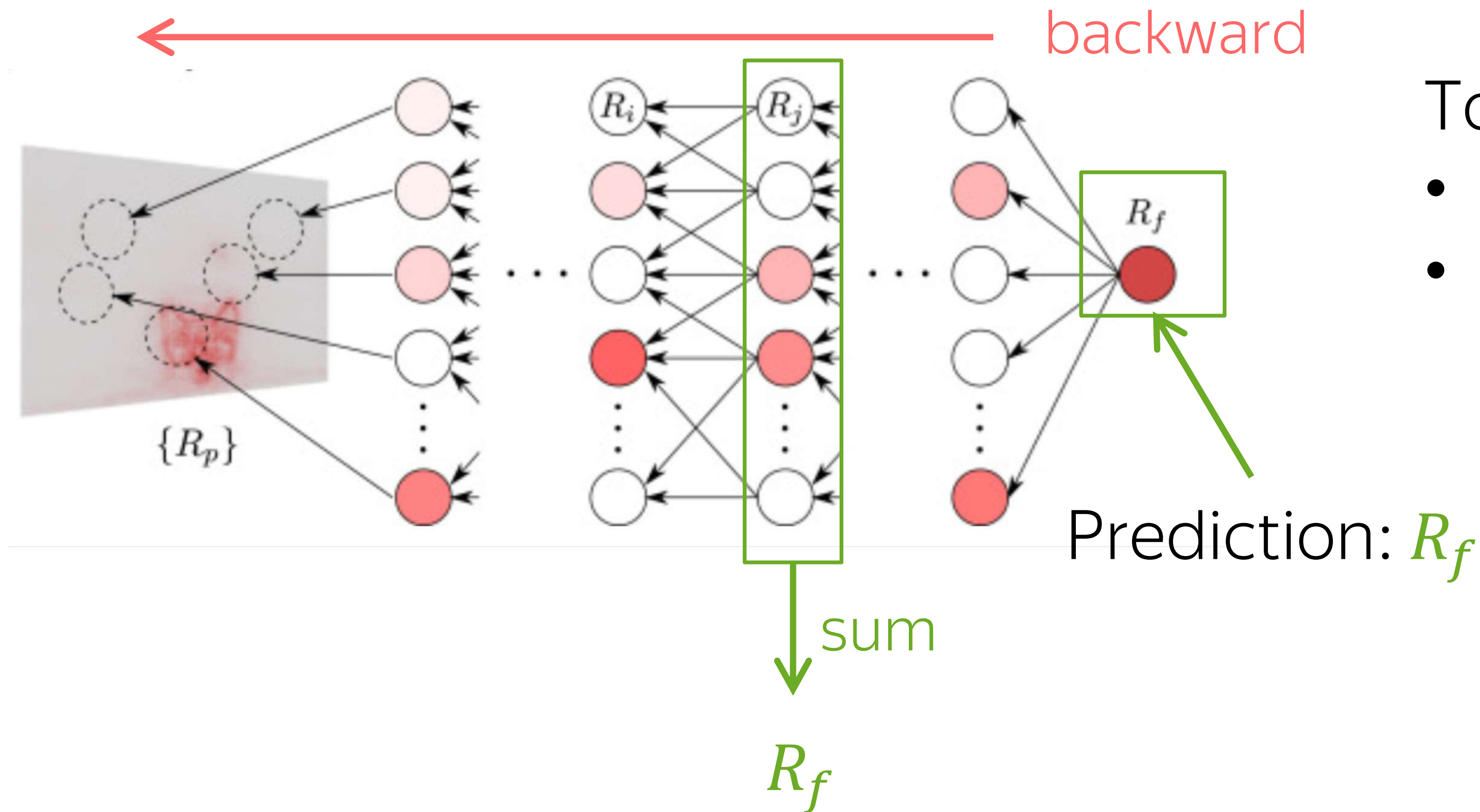
Total relevance is

- constant
- equals the prediction

Illustration from: <http://danshiebler.com/2017-04-16-deep-taylor-lrp/>

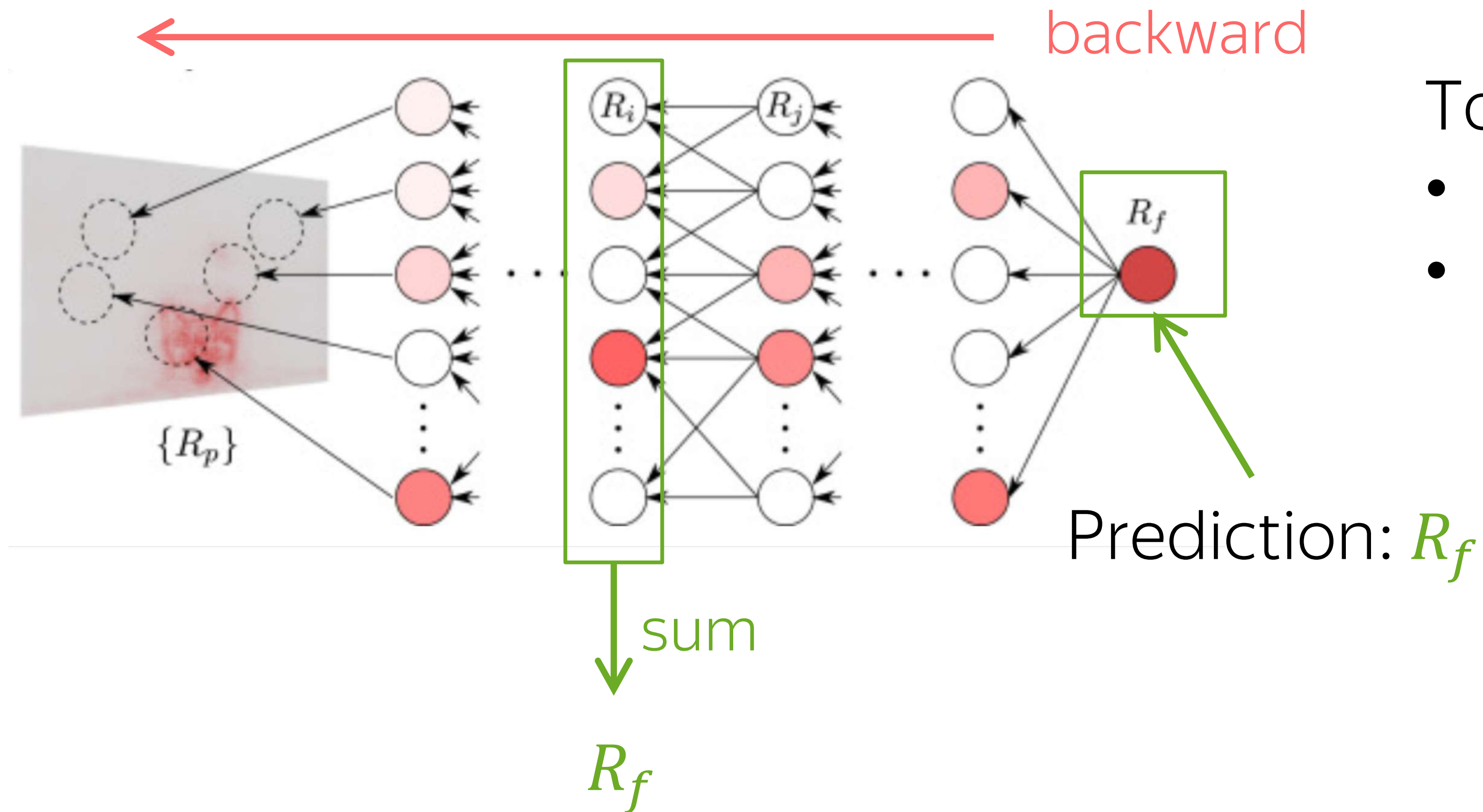


# Why LRP: Layerwise Conservation Principle



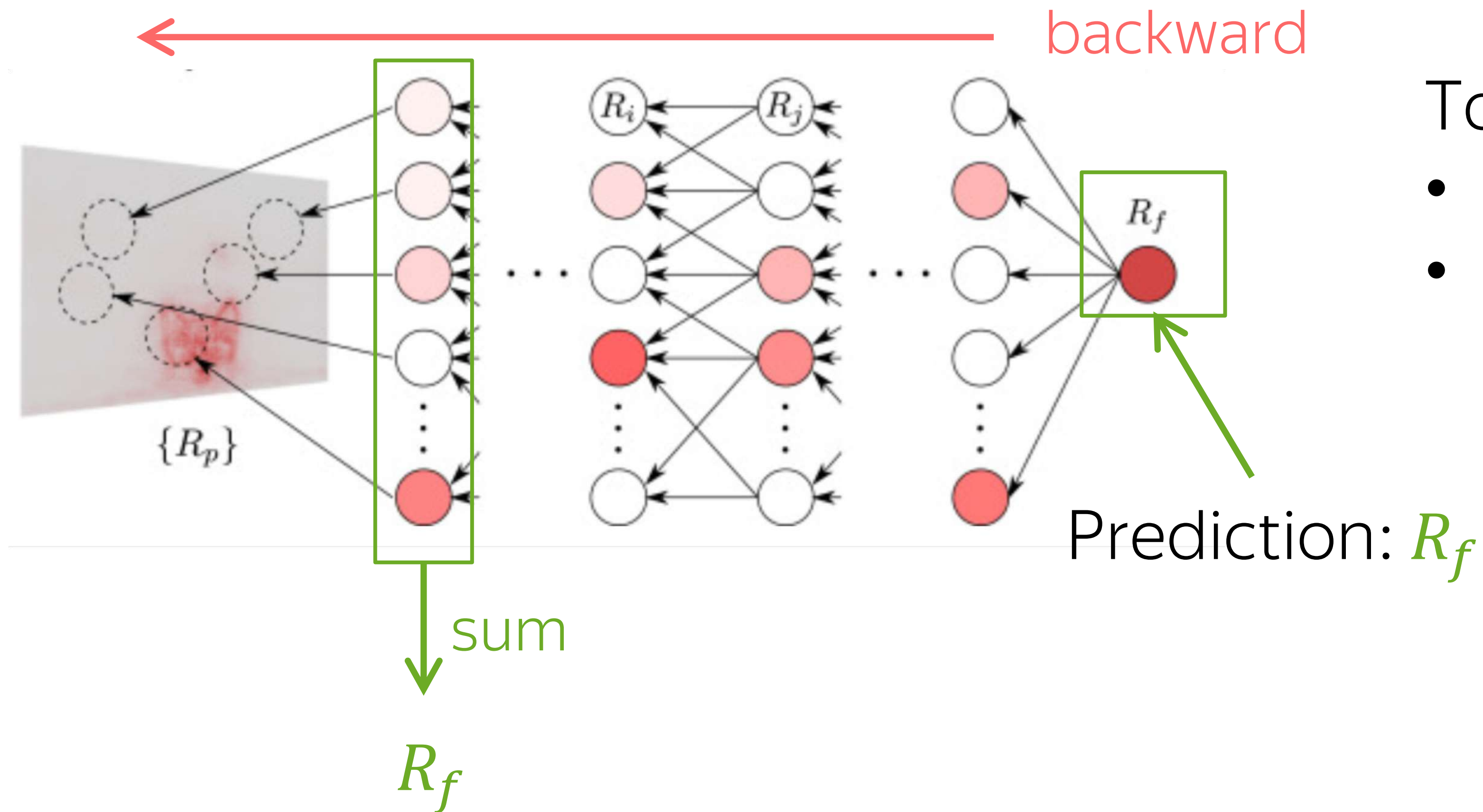
- Total relevance is
- constant
  - equals the prediction

# Why LRP: Layerwise Conservation Principle



- Total relevance is
- constant
  - equals the prediction

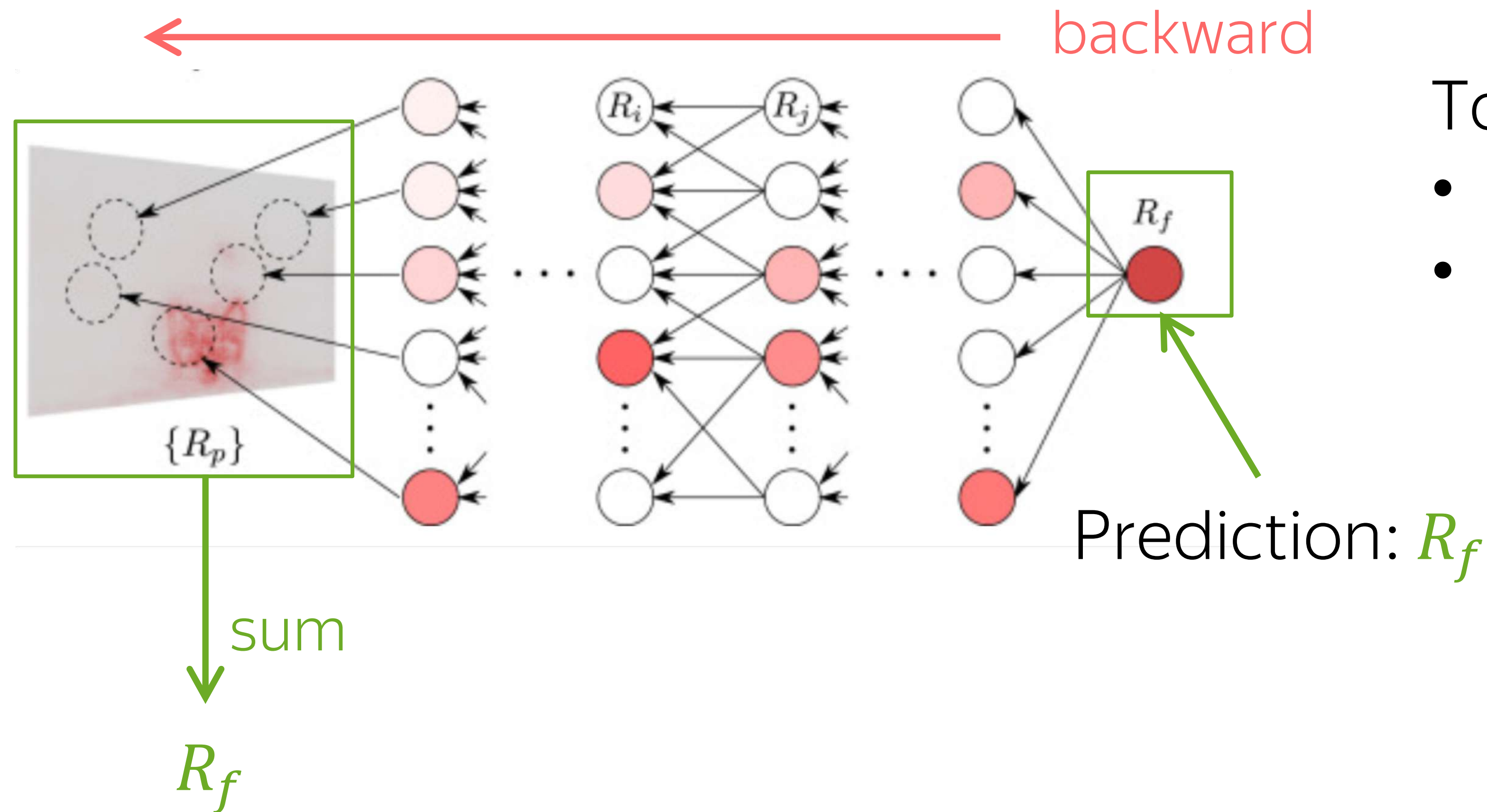
# Why LRP: Layerwise Conservation Principle



- Total relevance is
- constant
  - equals the prediction

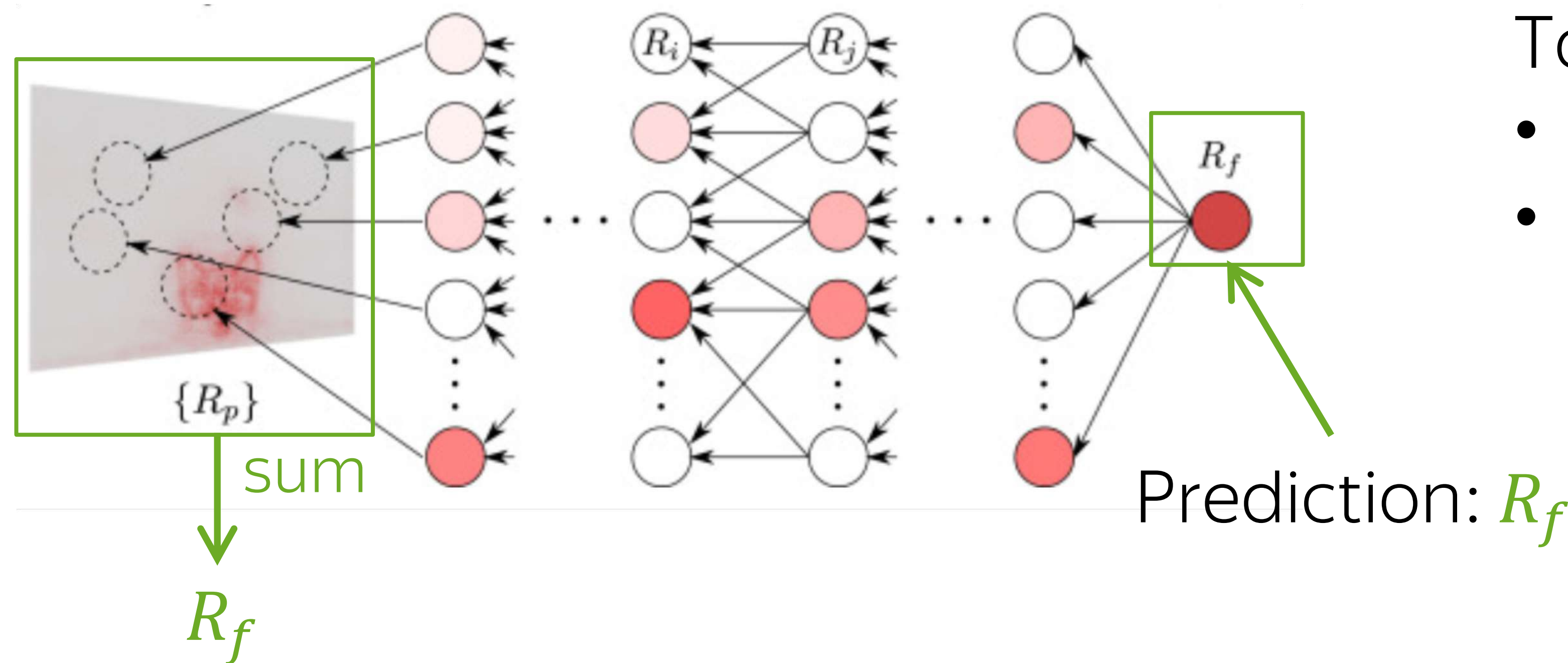


# Why LRP: Layerwise Conservation Principle



- Total relevance is
- constant
  - equals the prediction

# Why LRP: Layerwise Conservation Principle

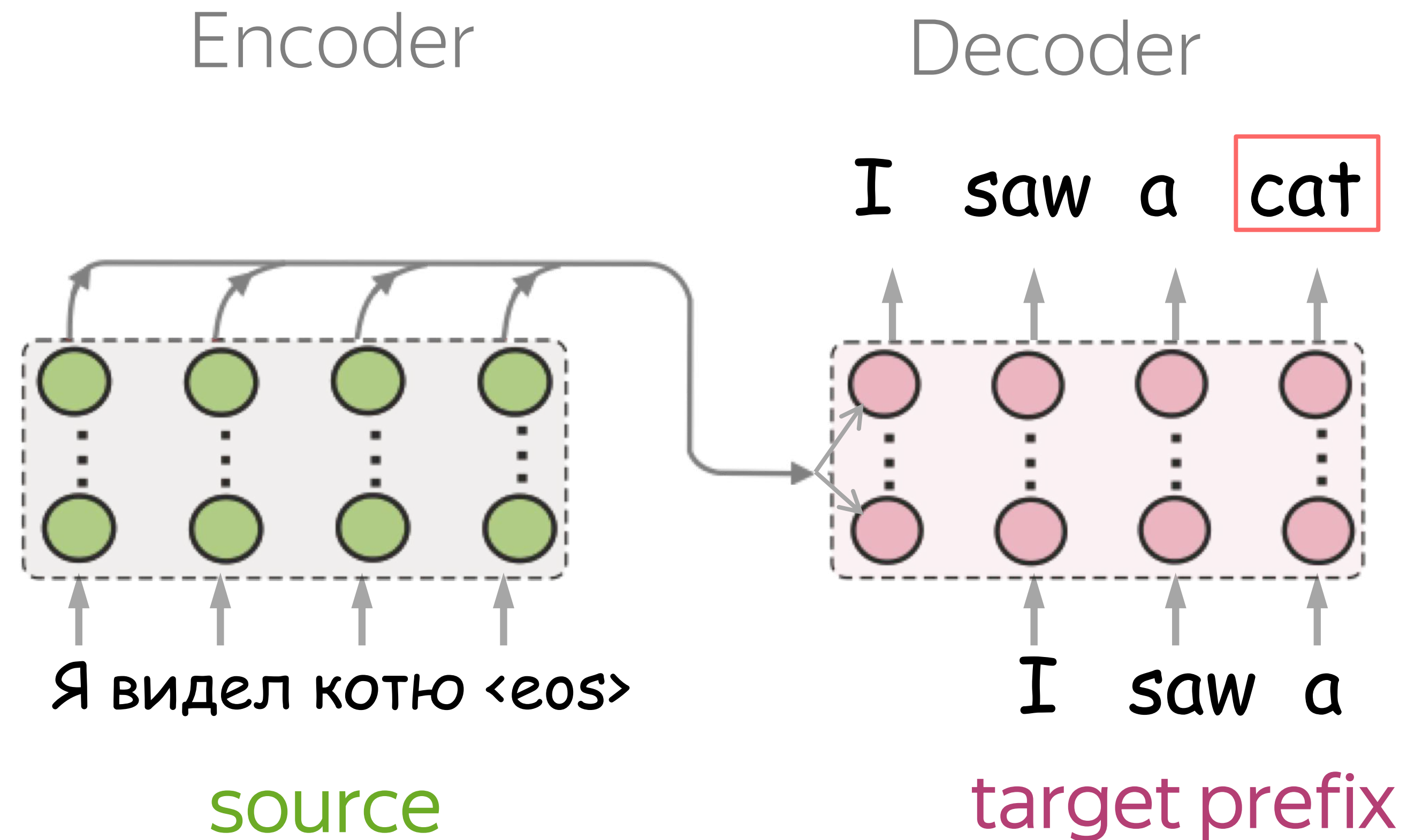


- Total relevance is
- constant
  - equals the prediction

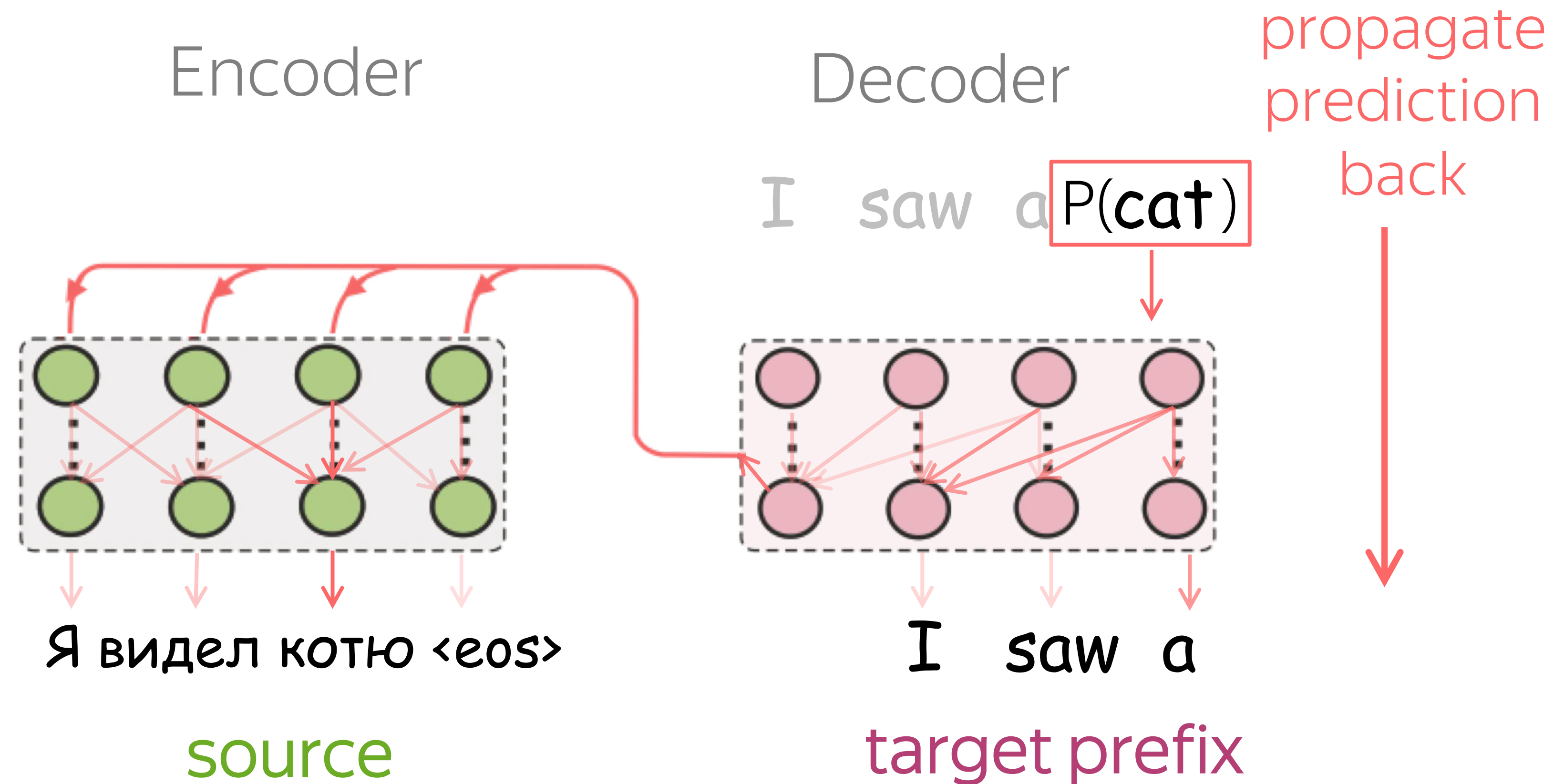
We can evaluate **relative** contribution of input elements!



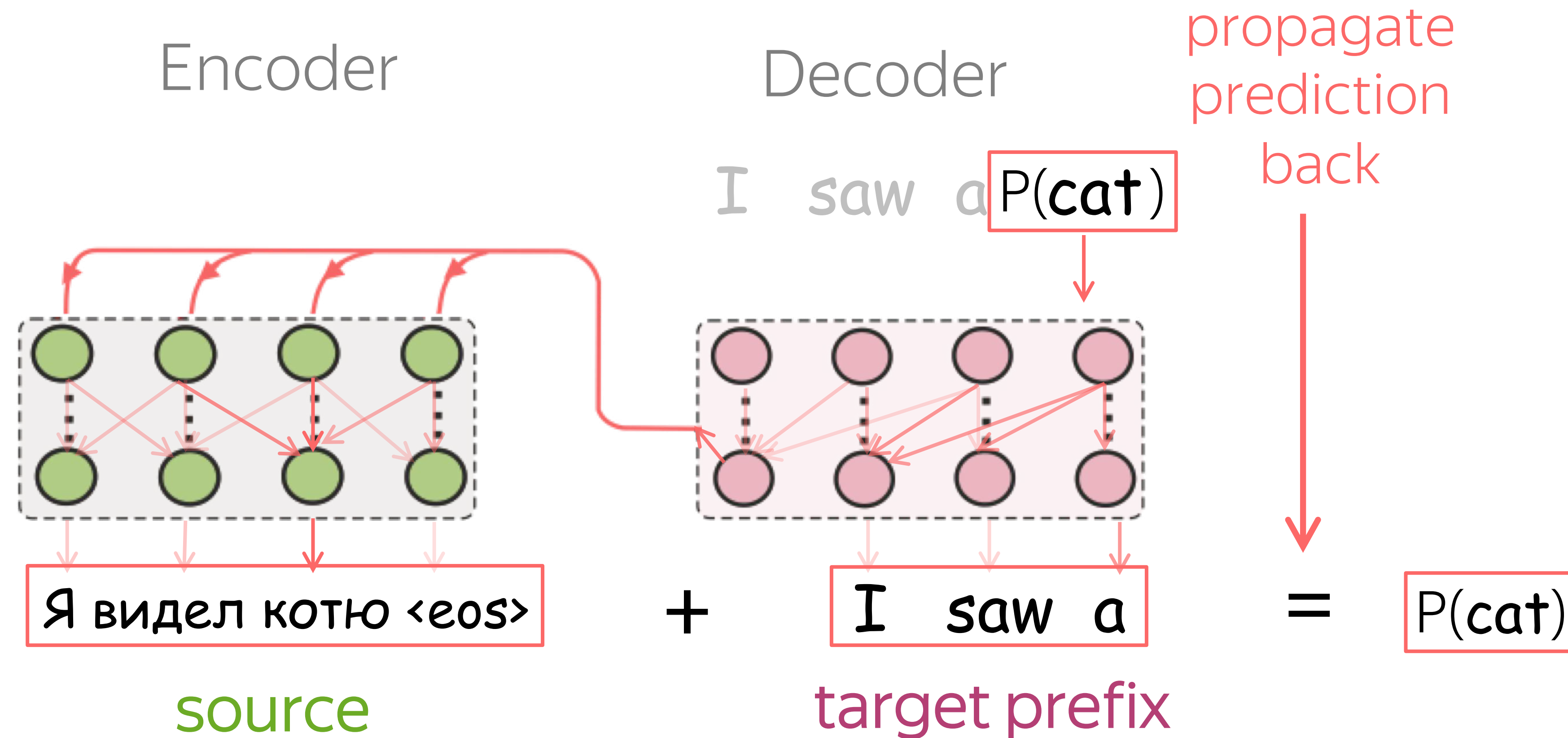
# Conservation Principle in NMT Models



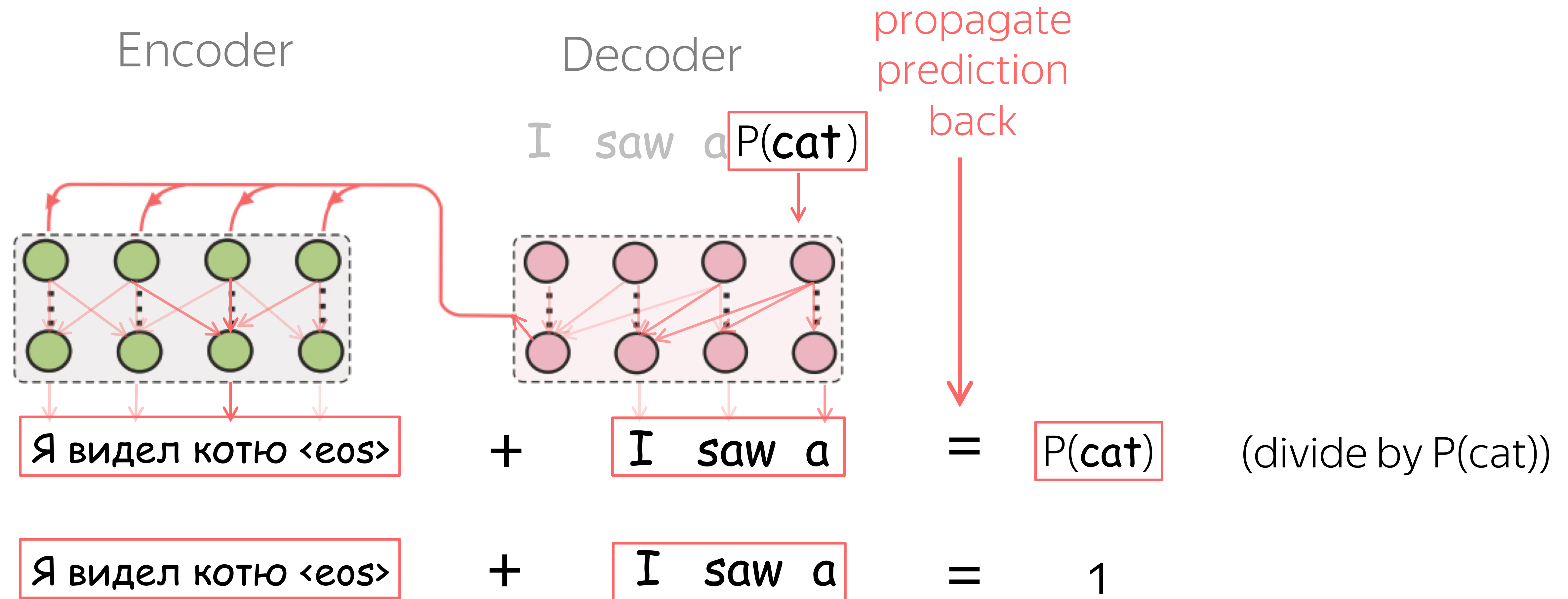
# Conservation Principle in NMT Models



# Conservation Principle in NMT Models



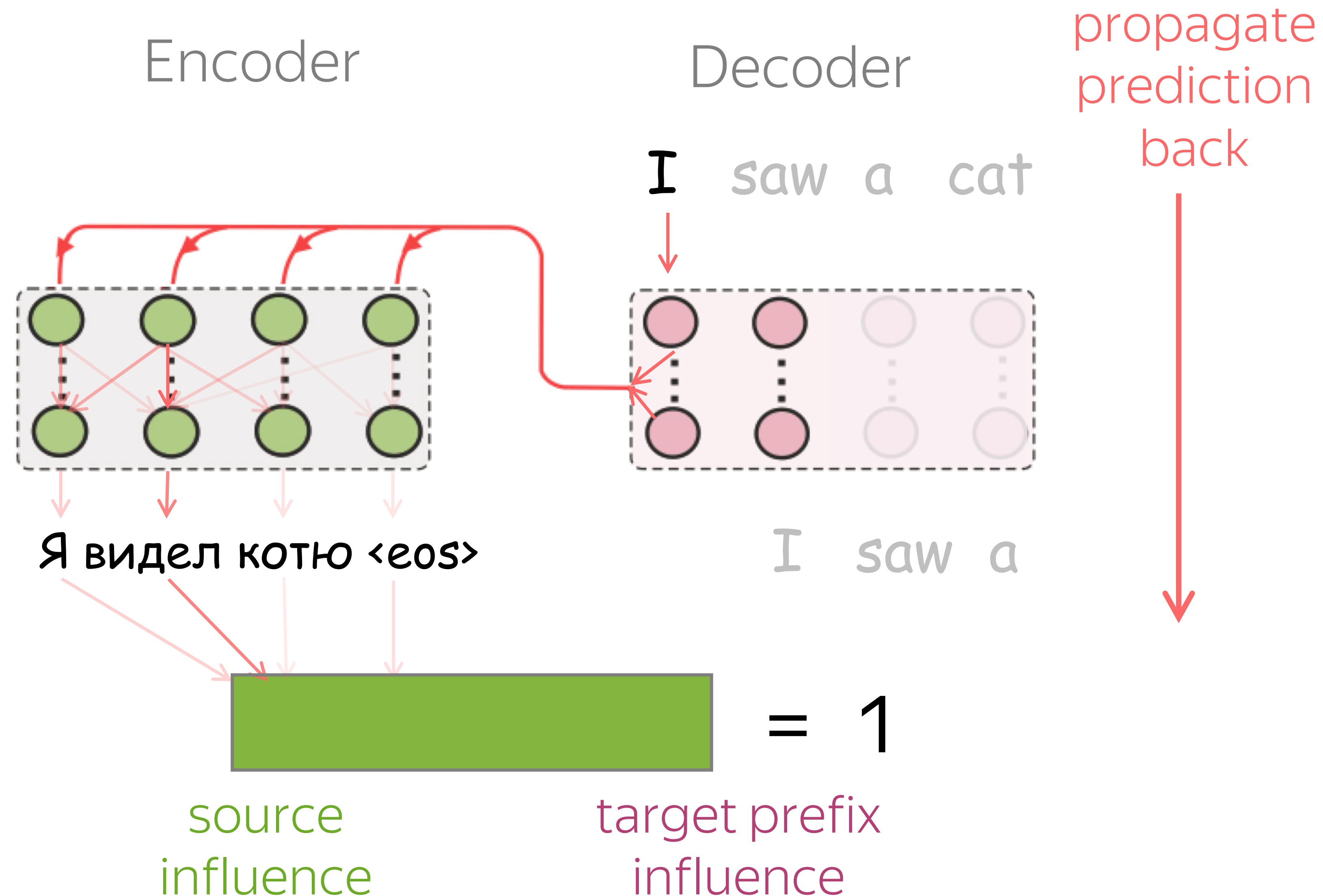
# Conservation Principle in NMT Models



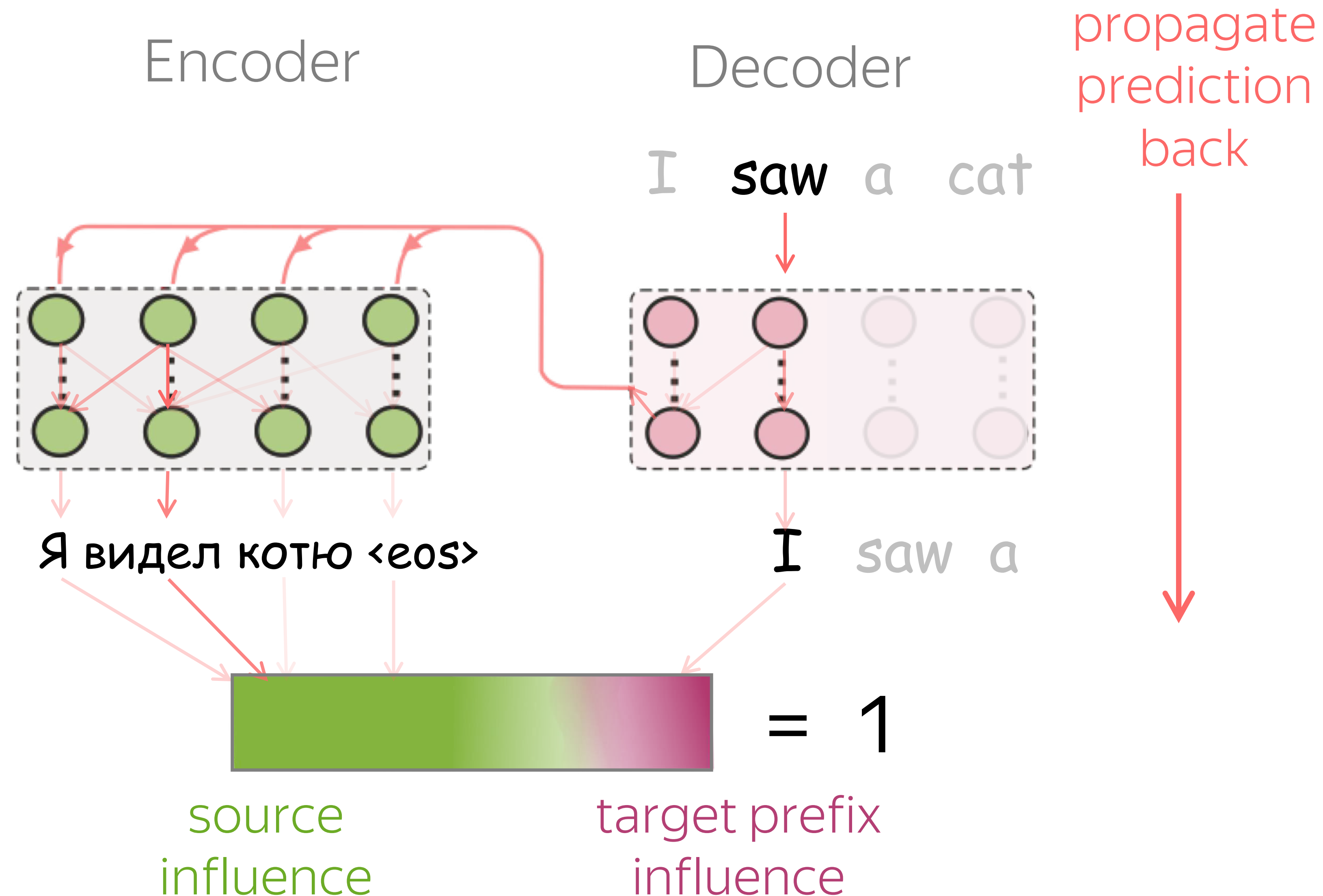
Relative contribution of source and target tokens



# Conservation Principle in NMT Models

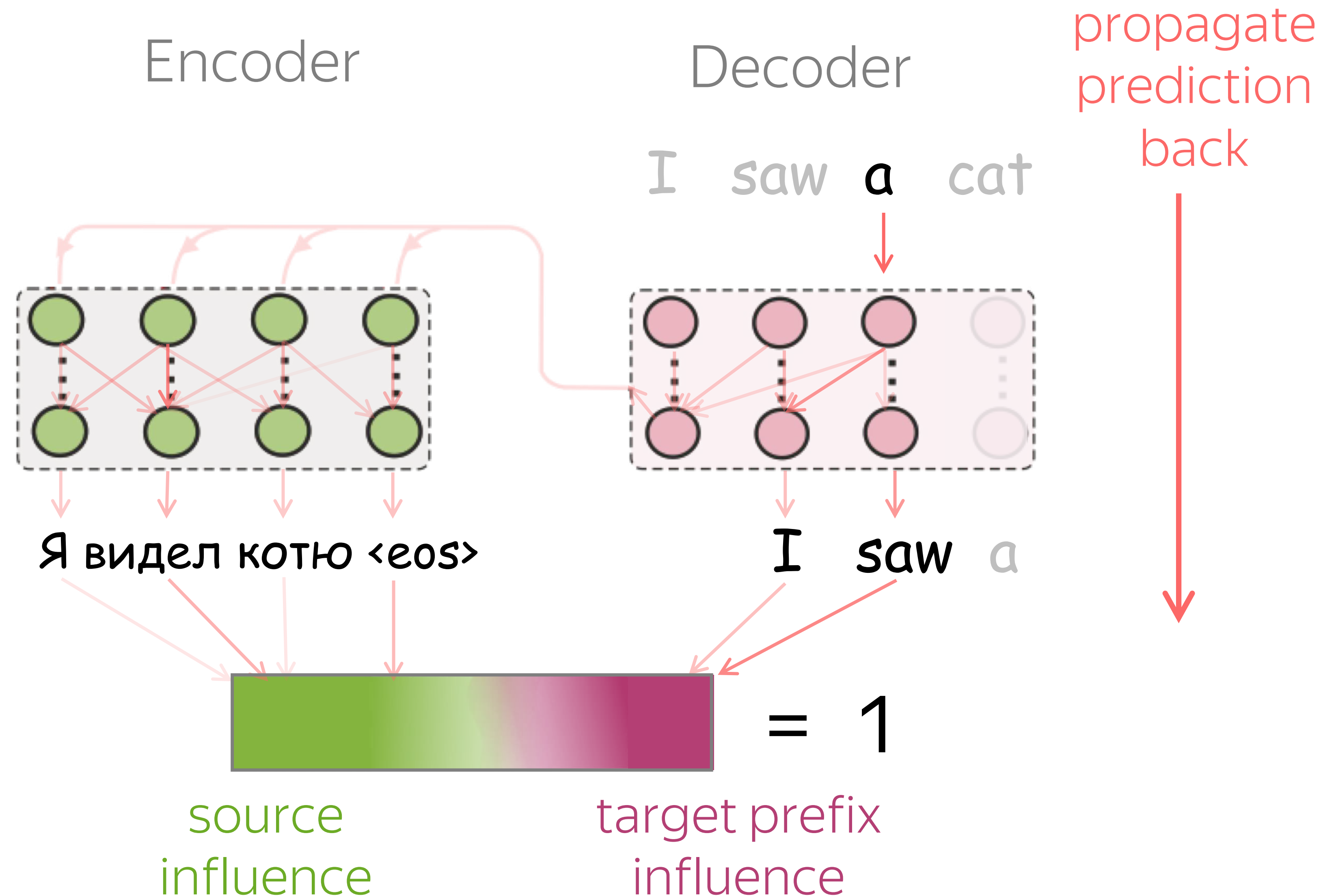


# Conservation Principle in NMT Models

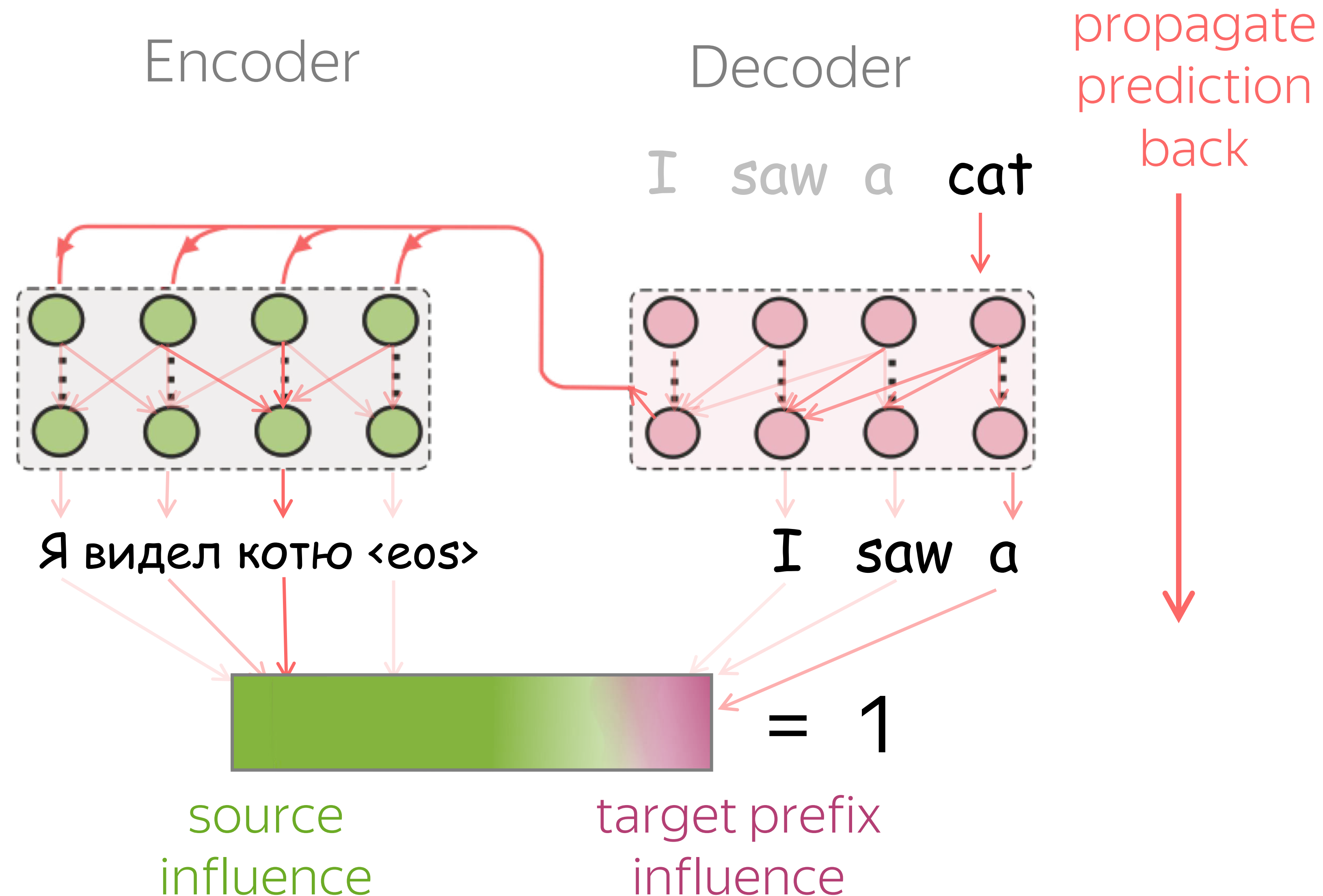




# Conservation Principle in NMT Models



# Conservation Principle in NMT Models

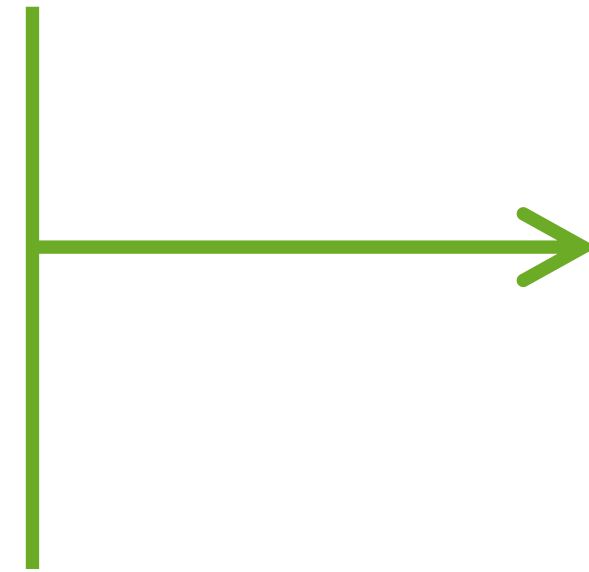


# LRP for Source and Target Contributions to NMT

- standard LRP does not support many operations (e.g., attention layer)

# LRP for Source and Target Contributions to NMT

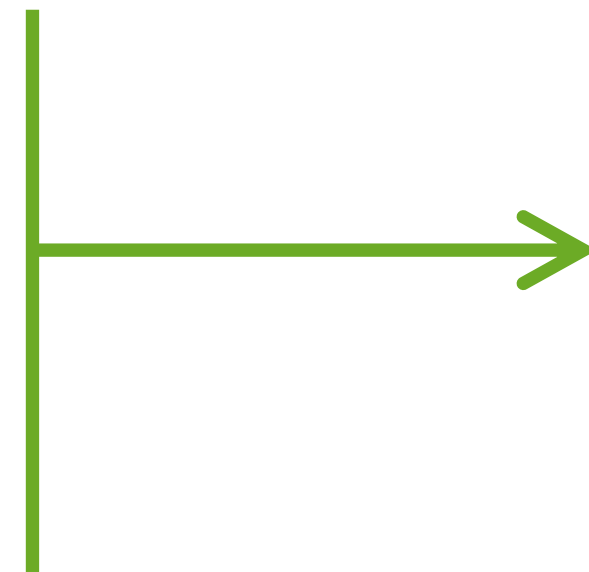
- standard LRP does not support many operations (e.g., attention layer)



We extend LRP to these layers

# LRP for Source and Target Contributions to NMT

- standard LRP does not support many operations (e.g., attention layer)

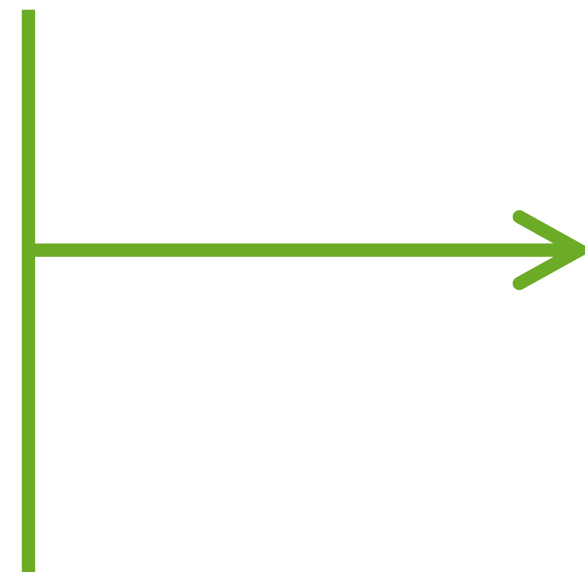


We extend LRP to these layers

- standard LRP has several variants:  $\varepsilon$ -,  $\alpha\beta$ -,  $\gamma$ -rules (differ in a way they redistribute relevance)

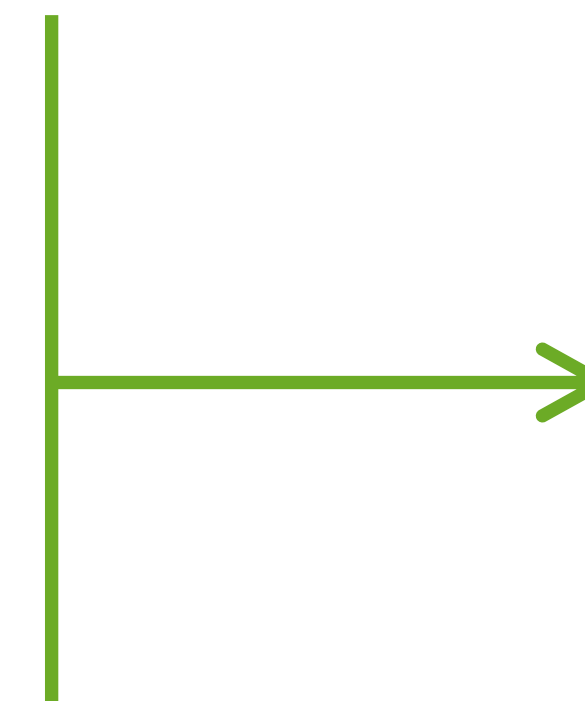
# LRP for Source and Target Contributions to NMT

- standard LRP does not support many operations (e.g., attention layer)



We extend LRP to these layers

- standard LRP has several variants:  $\varepsilon$ -,  $\alpha\beta$ -,  $\gamma$ -rules (differ in a way they redistribute relevance)

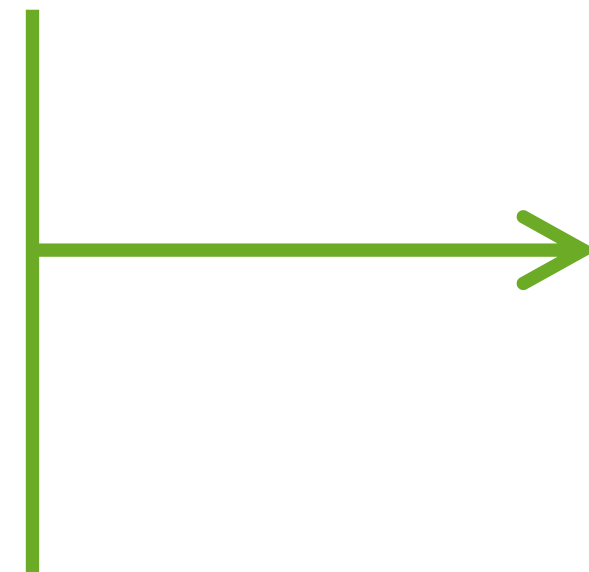


We use  $\alpha\beta$ -LRP: it keeps all contributions positive



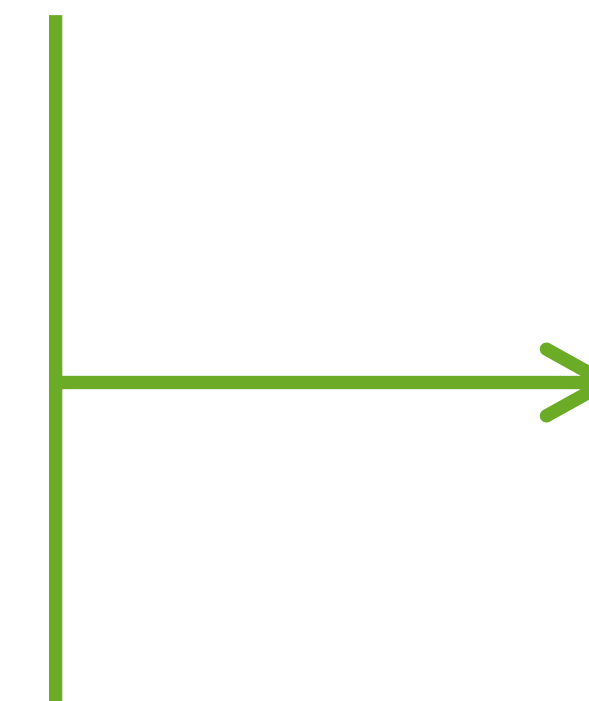
# LRP for Source and Target Contributions to NMT

- standard LRP does not support many operations (e.g., attention layer)



We extend LRP to these layers

- standard LRP has several variants:  $\varepsilon$ -,  $\alpha\beta$ -,  $\gamma$ -rules (differ in a way they redistribute relevance)



We use  $\alpha\beta$ -LRP: it keeps all contributions positive

More details are in the paper!

# What We Do

LRP:

- can be inaccurate for small contributions
- relevancies may differ little

# What We Do

LRP:

- can be inaccurate for small contributions
- relevancies may differ little



We talk  
about

- general patterns  
but not individual examples (i.e.,  
we average over a dataset)

# What We Do

LRP:

- can be inaccurate for small contributions
- relevancies may differ little



We talk  
about

- general patterns  
but **not** individual examples (i.e.,  
we average over a dataset)
- how these patterns change  
(e.g., across models, datasets,  
training stages, etc)  
but **not** about absolute values of  
contributions

# What is going to happen:

## The Trade-Off Between Source and Target

- Our Approach: (a version of) LRP
- Experiments
  - Getting Acquainted
  - Reference, Model and Random Prefixes
  - Exposure Bias and Source Contribution
  - Varying the Amount of Data
  - Training Stages

(A Bit of) the Training Process (work in progress)



# What is going to happen:

## The Trade-Off Between Source and Target

- Our Approach: (a version of) LRP
- Experiments
  - Getting Acquainted
  - Reference, Model and Random Prefixes
  - Exposure Bias and Source Contribution
  - Varying the Amount of Data
  - Training Stages

Compare patterns for  
different models

(A Bit of) the Training Process (work in progress)

# What is going to happen:

## The Trade-Off Between Source and Target

- Our Approach: (a version of) LRP
- Experiments
  - Getting Acquainted
  - Reference, Model and Random Prefixes
  - Exposure Bias and Source Contribution
  - Varying the Amount of Data
  - Training Stages

Compare patterns for  
different models

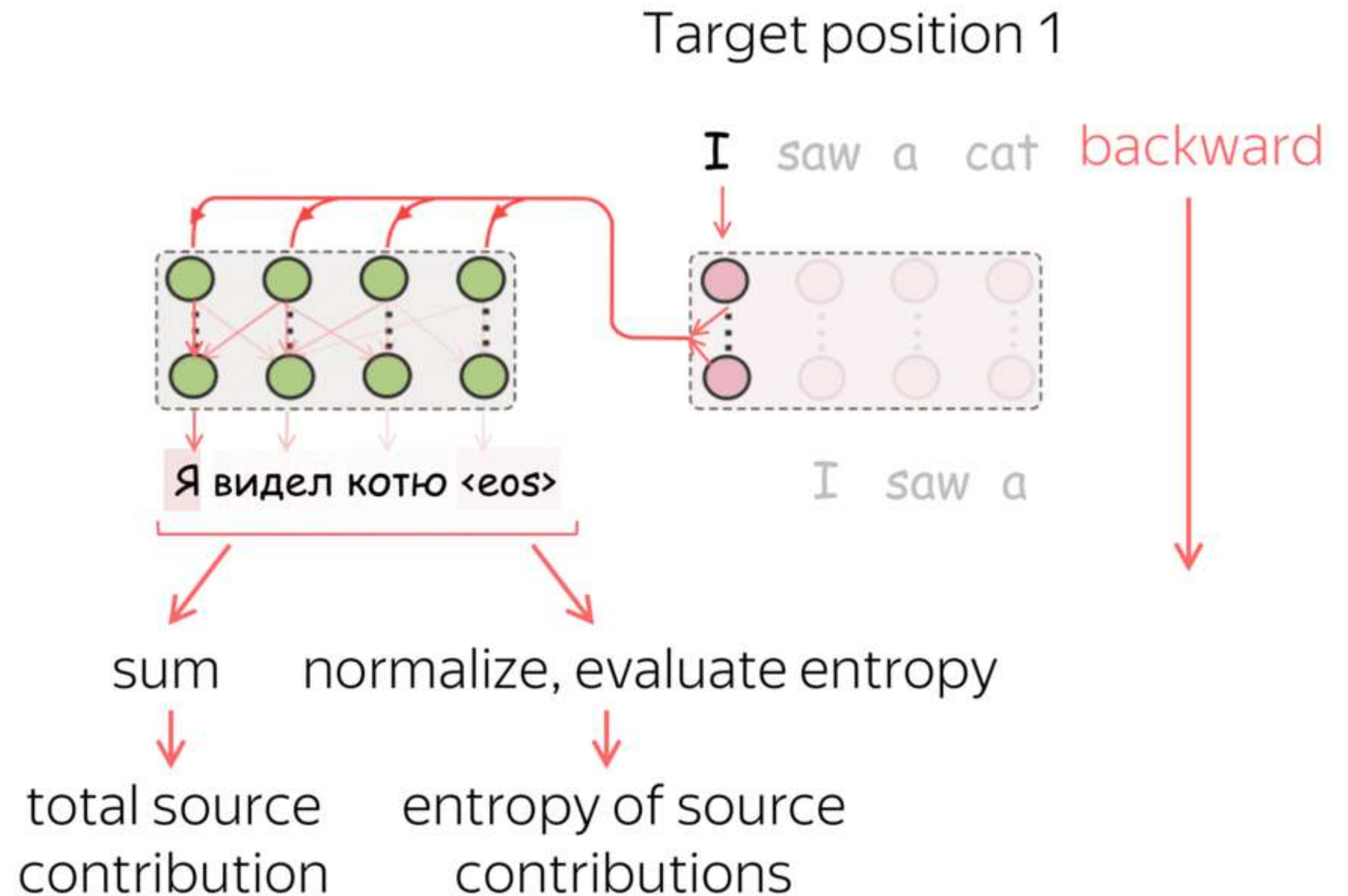
(A Bit of) the Training Process (work in progress)

# Experiments: Getting Acquainted



# We look at: total contribution and entropy

- separately for each target position
- total contribution or entropy of contributions





# We look at: average over a dataset

source

target

"I"	"saw"	"cat"							
Я	видел	котю .	<eos>	→	I	saw	a	cat .	<eos>
					1	0.8	0.3	0.9	0.5 0.4
"Cat"	"was"	"cute"							
Котя	был	милый .	<eos>	→	The	cat	was	cute .	<eos>
					1	0.9	0.7	0.8	0.4 0.6
"I"	"fed"	"cat"							
Я	покармил	котю .	<eos>	→	I	fed	the	cat .	<eos>
					1	0.7	0.1	0.8	0.6 0.3
...	...	...	... <eos>	→	...	...	...	...	... <eos>
					...	...	...	...	...
same sentence length					same sentence length				

# We look at: average over a dataset

source

target

E.g., source contributions  
at each target step

"I"	"saw"	"cat"									
Я	видел	котю .	<eos>	→	I	saw	a	cat	.	<eos>	✓
					1	0.8	0.3	0.9	0.5	0.4	
"Cat"	"was"	"cute"									
Котя	был	милый .	<eos>	→	The	cat	was	cute	.	<eos>	
					1	0.9	0.7	0.8	0.4	0.6	
"I"	"fed"	"cat"									
Я	покормил	котю .	<eos>	→	I	fed	the	cat	.	<eos>	
					1	0.7	0.1	0.8	0.6	0.3	
...	...	...	... <eos>	→	...	...	...	...	...	<eos>	
					...	...	...	...	...	...	
same sentence length					same sentence length						

# We look at: average over a dataset

source

target

E.g., source contributions  
at each target step

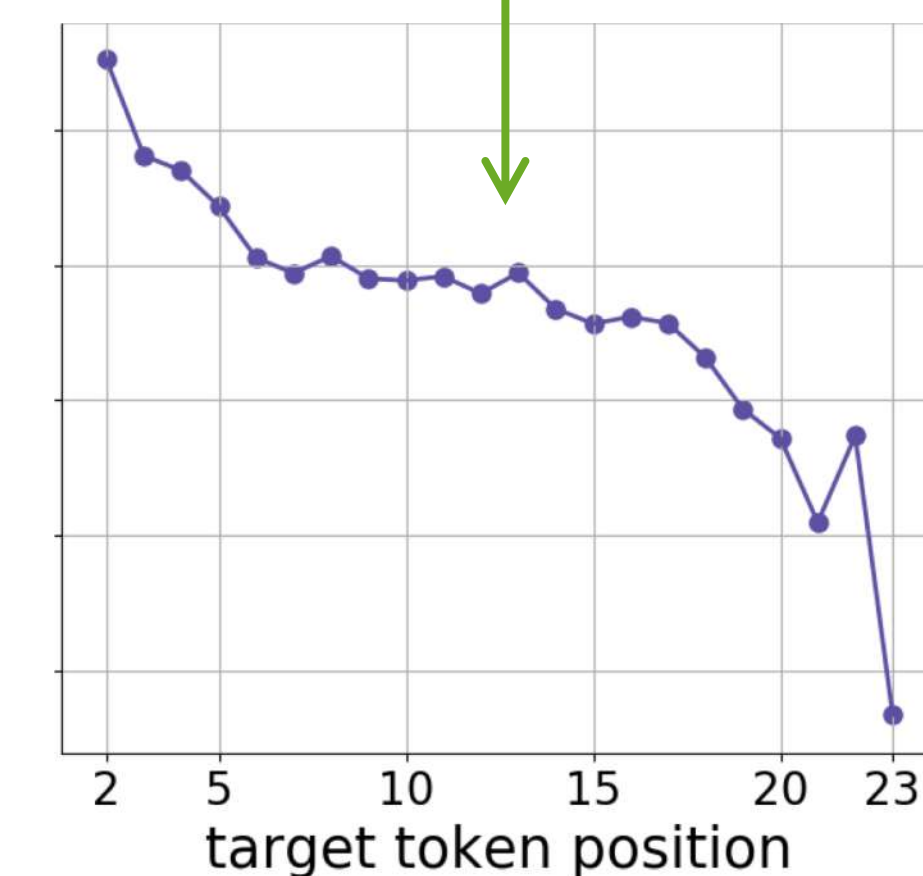
"I"	"saw"	"cat"									
Я	видел	котю	.	<eos>	→	I	saw	a	cat	.	<eos>
						1	0.8	0.3	0.9	0.5	0.4
"Cat"	"was"	"cute"									
Котя	был	милый	.	<eos>	→	The	cat	was	cute	.	<eos>
						1	0.9	0.7	0.8	0.4	0.6
"I"	"fed"	"cat"									
Я	покармил	котю	.	<eos>	→	I	fed	the	cat	.	<eos>
						1	0.7	0.1	0.8	0.6	0.3
...	...	...	...	<eos>	→	...	...	...	...	...	<eos>
						...	...	...	...	...	...

same sentence length

same sentence length

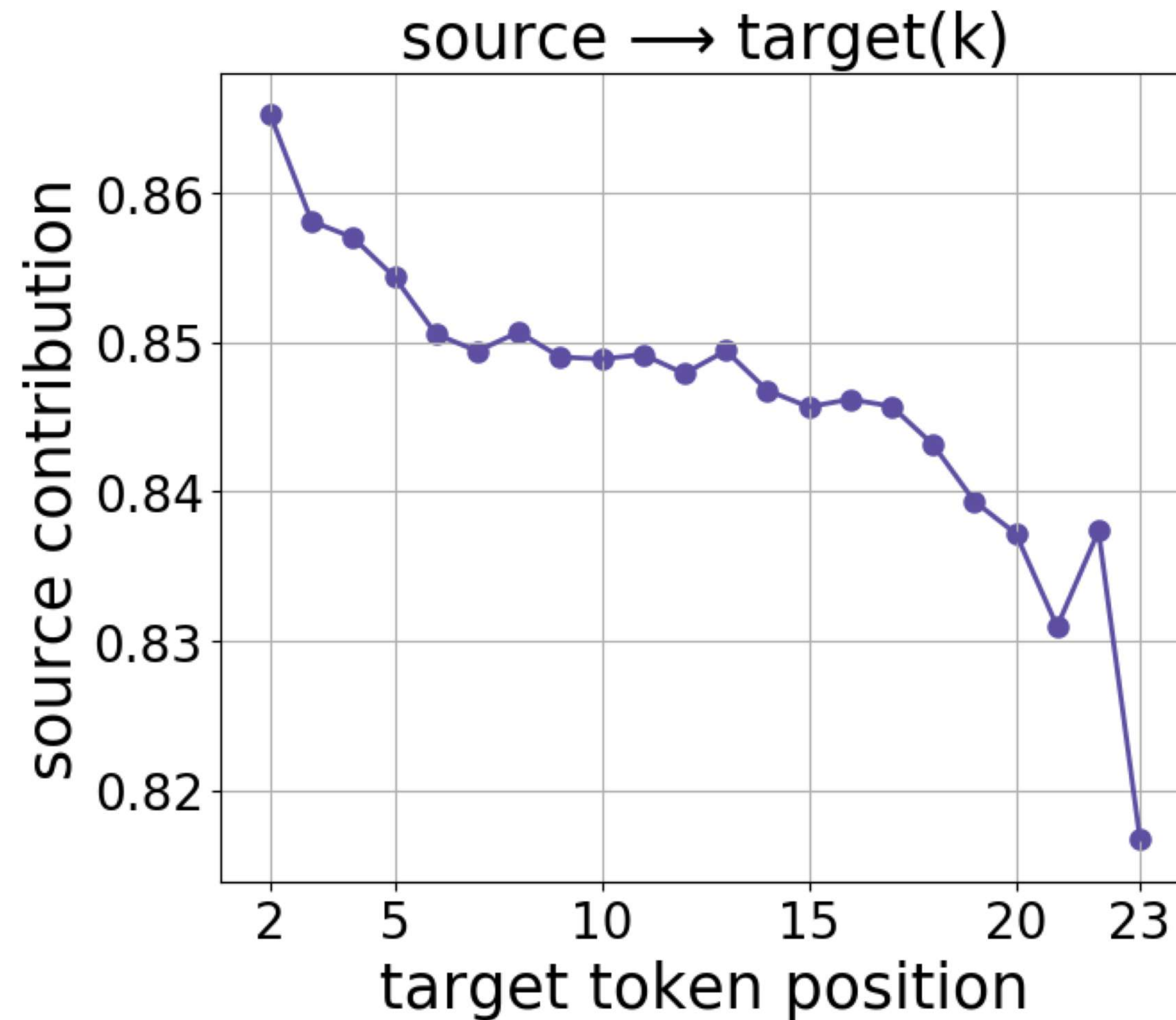
Average over an  
evaluation set

e.g., 1 0.9 0.87 ... ..



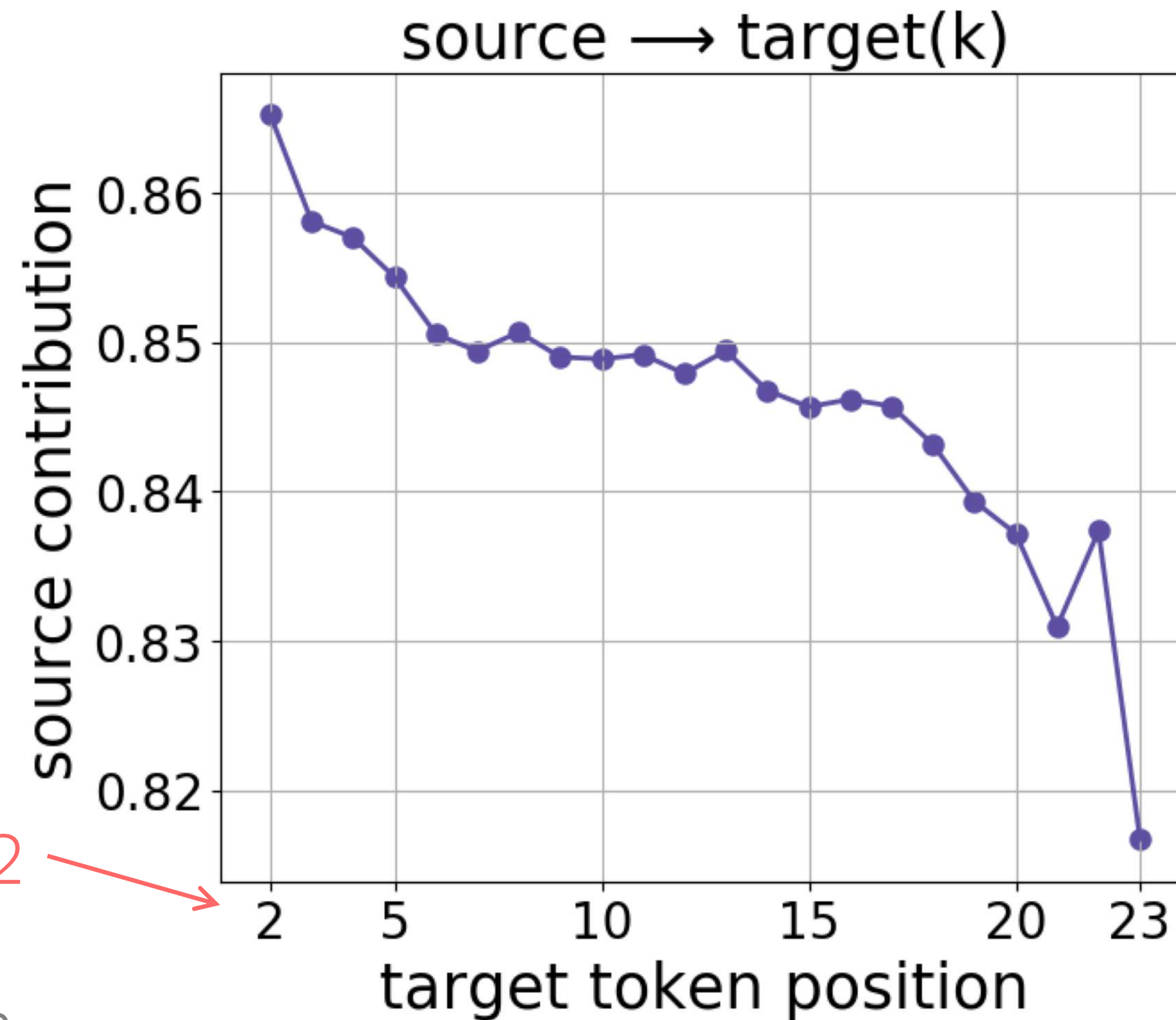
We see: general pattern

# Source Contribution to Different Target Positions





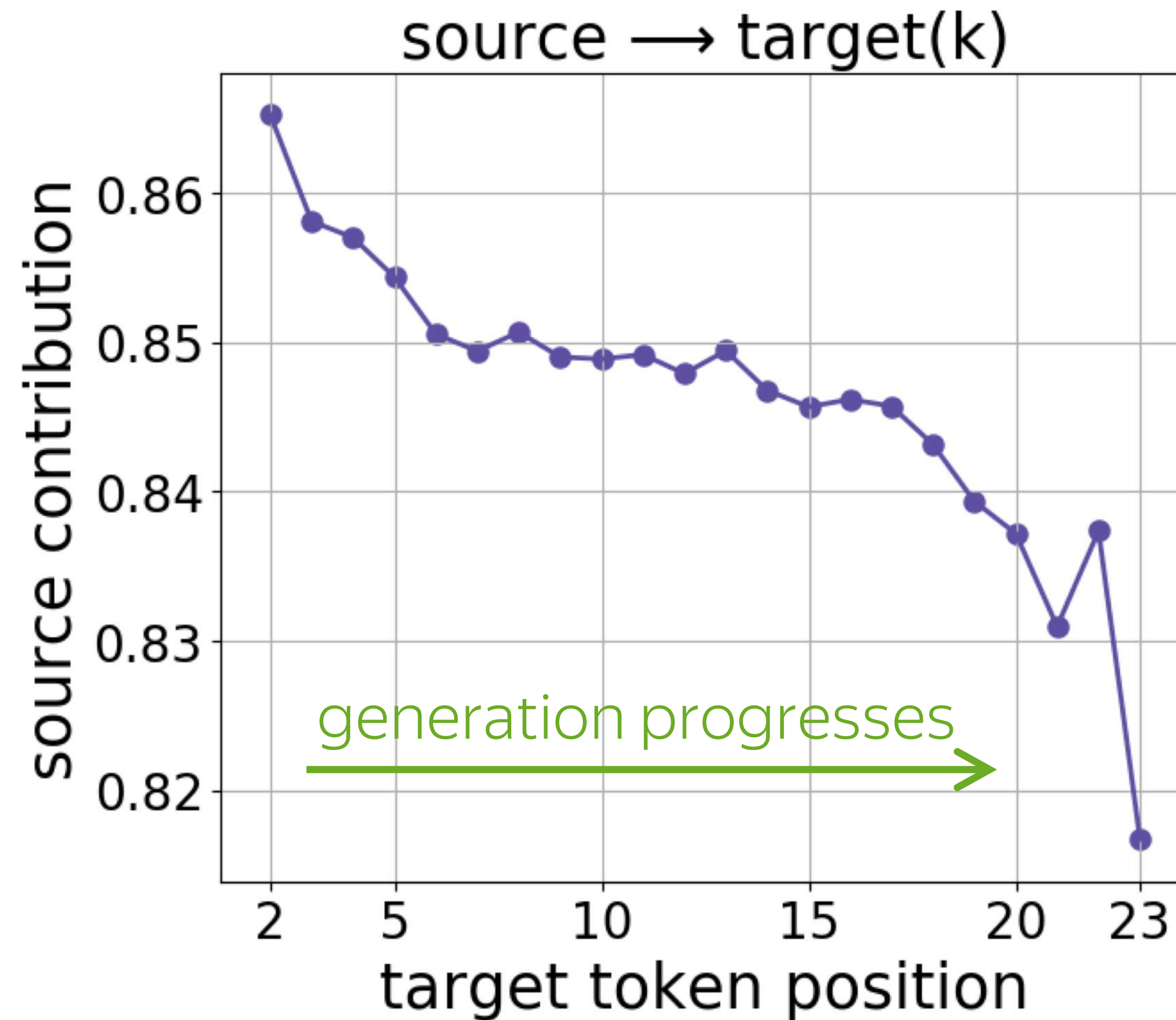
# Source Contribution to Different Target Positions



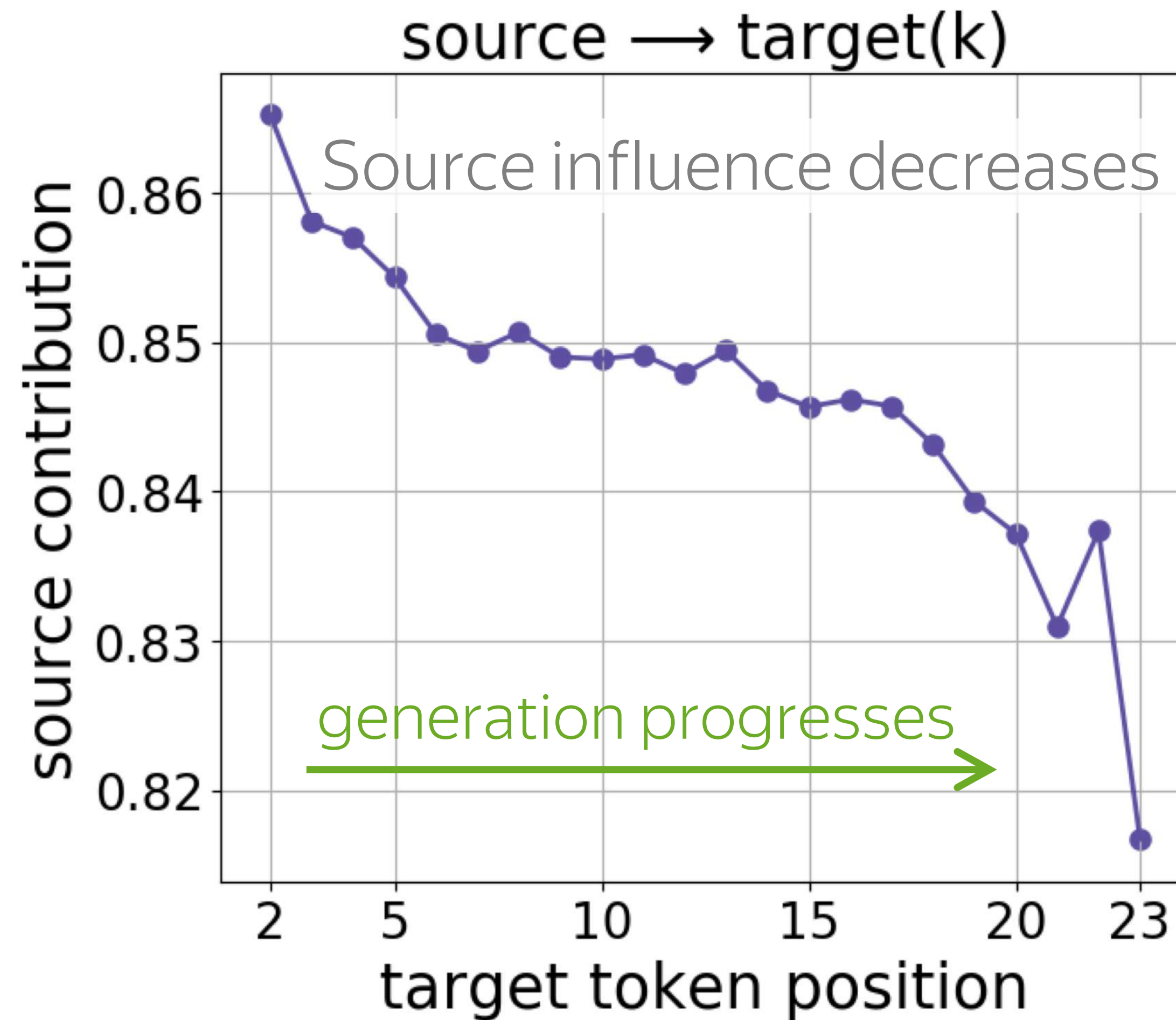
start from position 2  
(for the first token,  
source contribution  
is always 1)



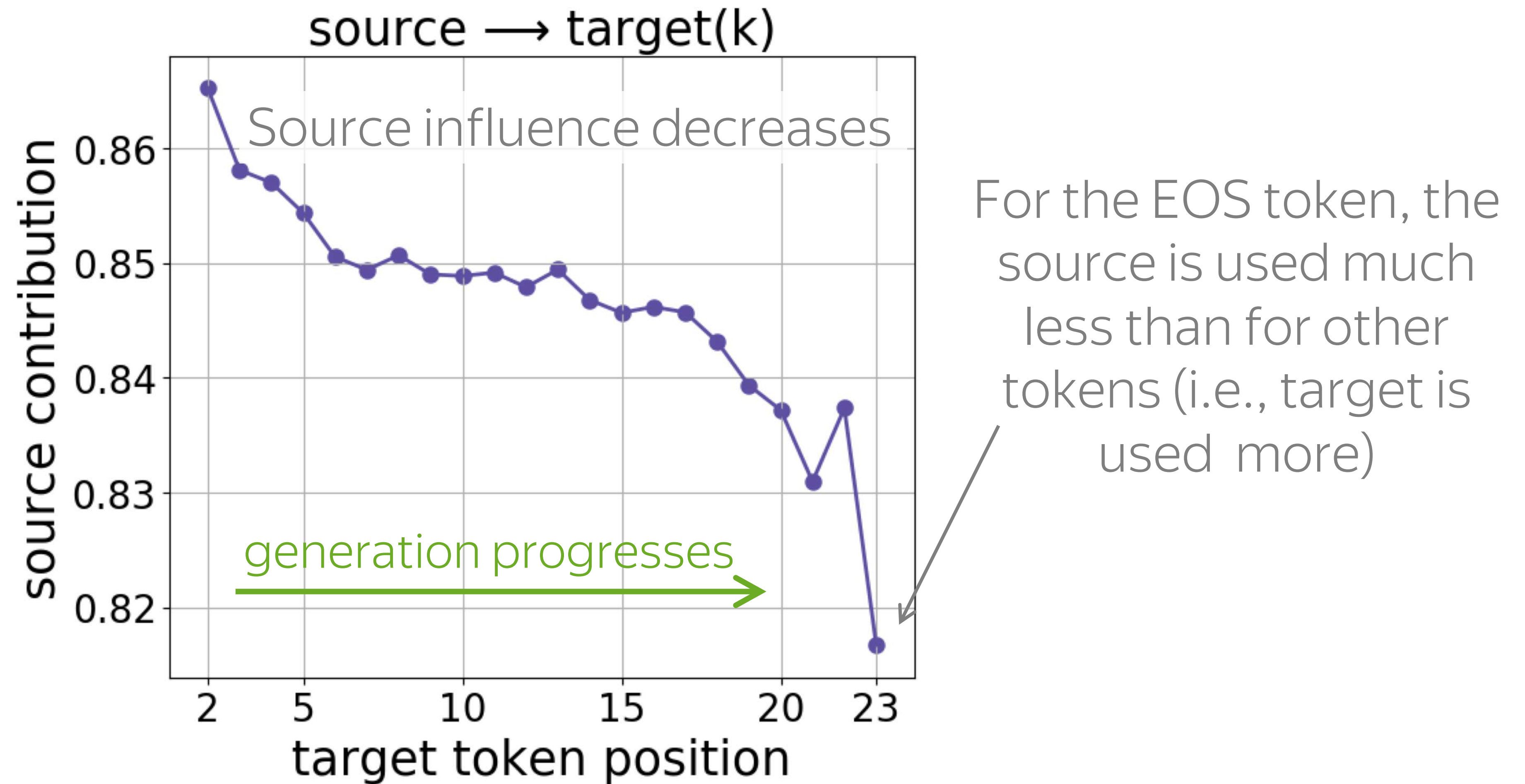
# Source Contribution to Different Target Positions



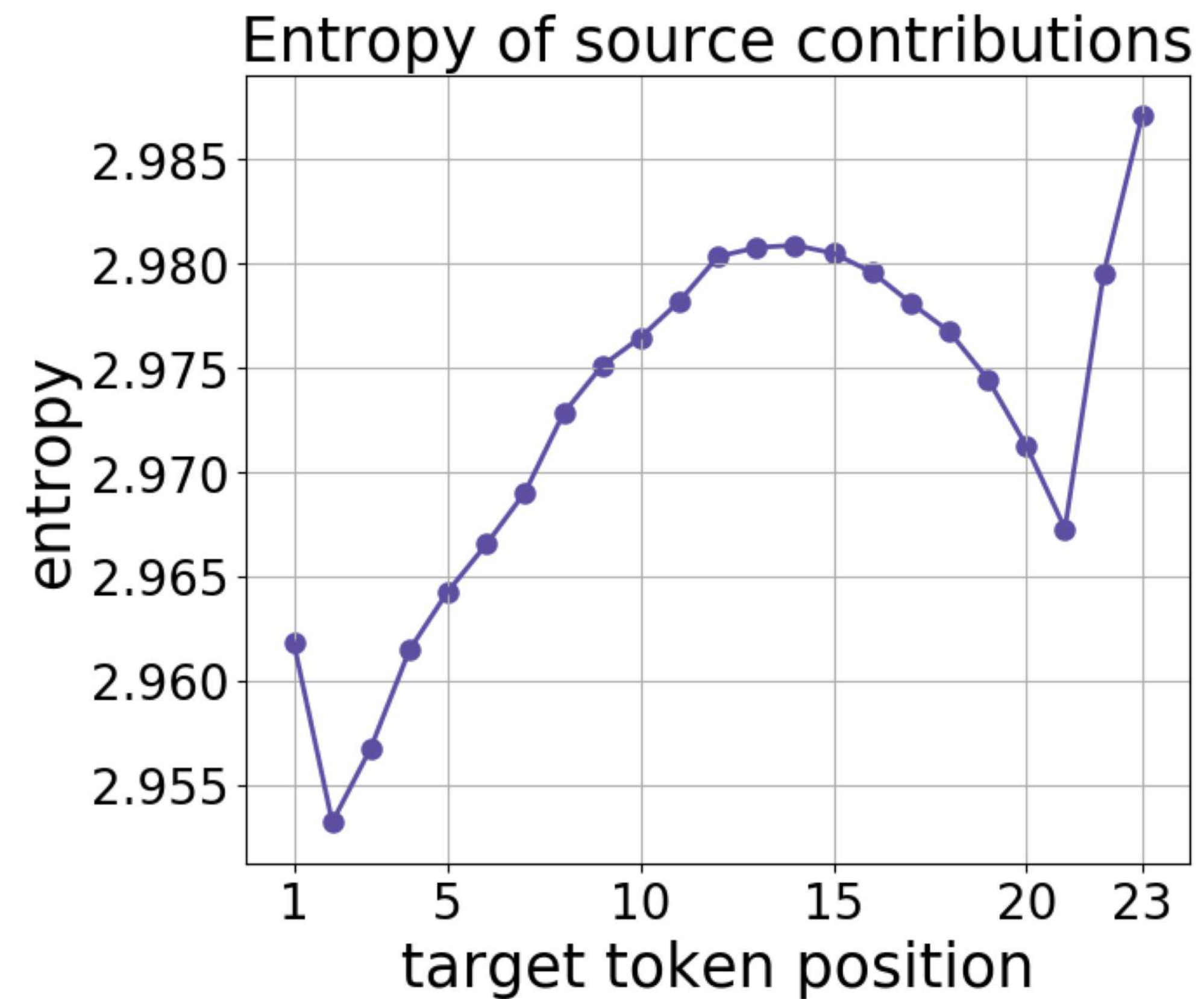
# Source Contribution to Different Target Positions



# Source Contribution to Different Target Positions

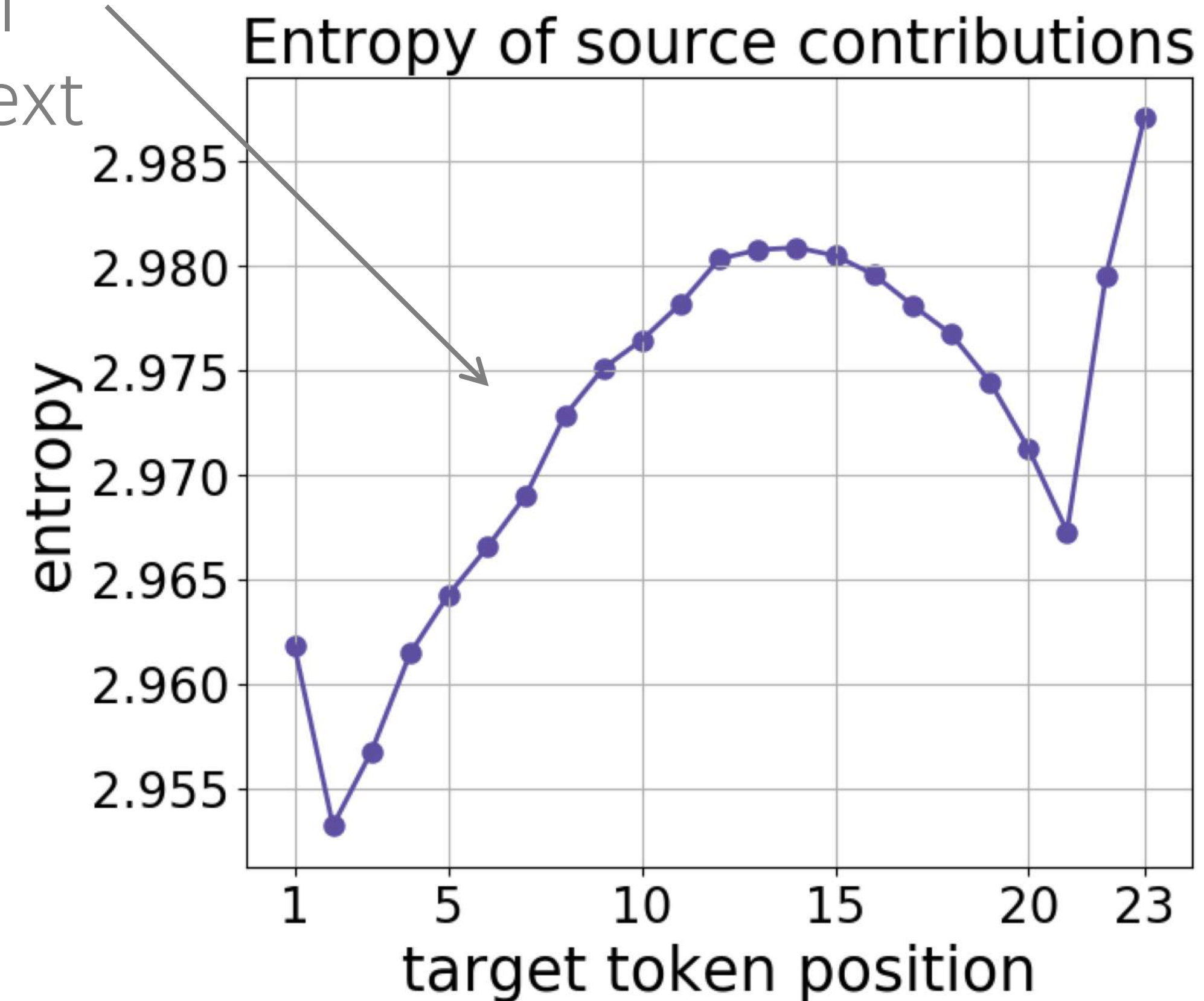


# Entropy of Source Contributions



# Entropy of Source Contributions

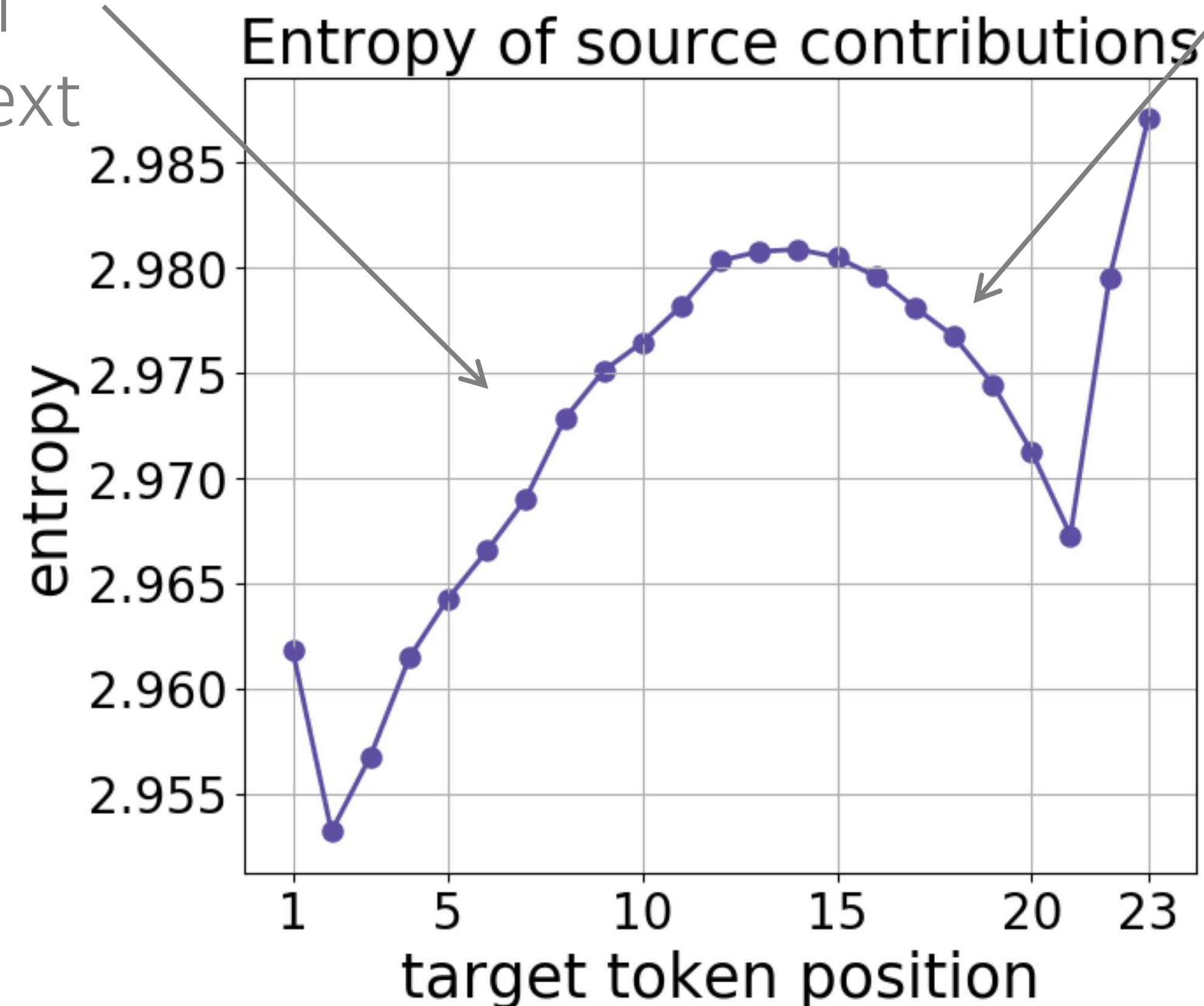
First, as generation progresses, the model relies on a broader context (entropy increases)





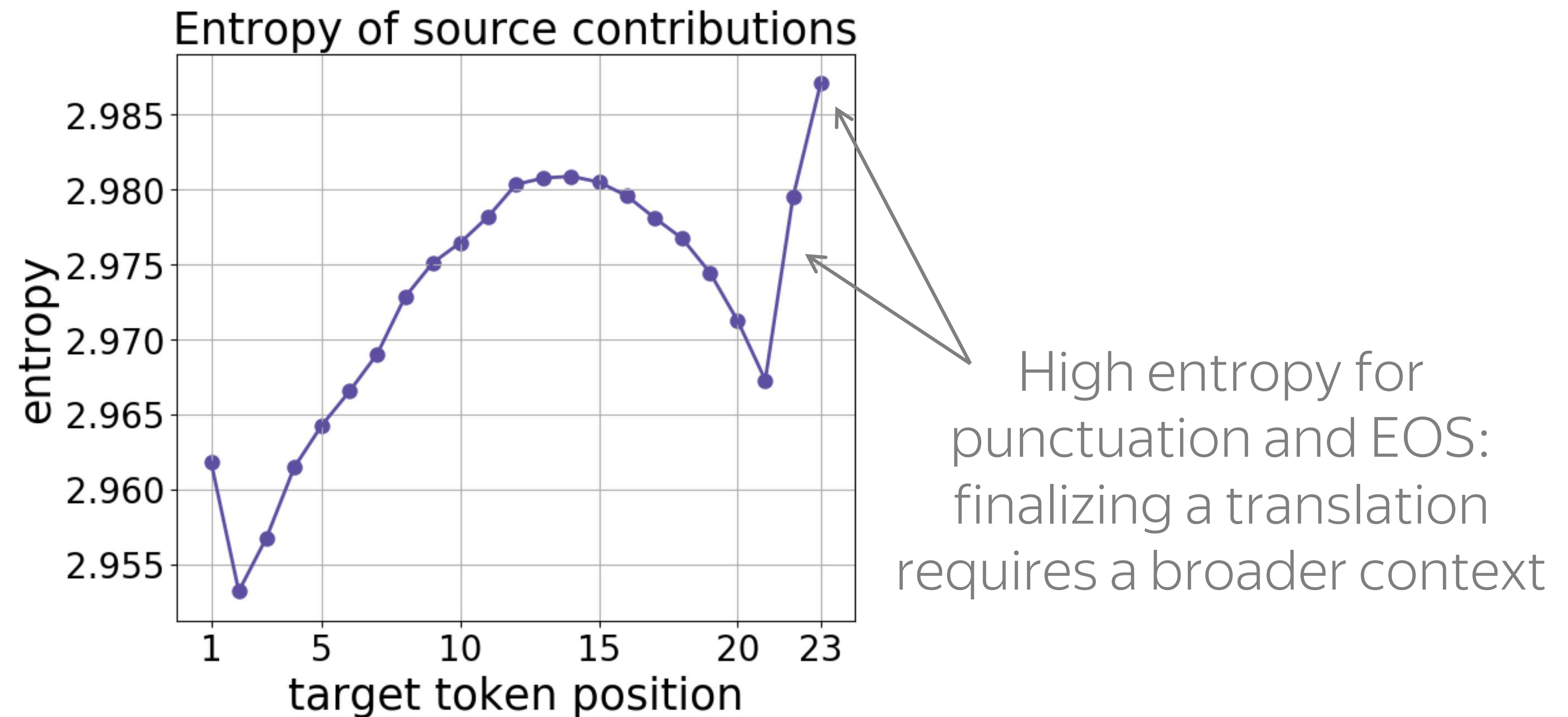
# Entropy of Source Contributions

First, as generation progresses, the model relies on a broader context (entropy increases)



Then, for the last part of a translation, it becomes more focused (entropy decreases)

# Entropy of Source Contributions



# Summary

During generation,

- source influence decreases (i.e., prefix influence increases)
- entropy of source contributions goes up till the half of the translation, then down

# What is going to happen:

## The Trade-Off Between Source and Target

- Our Approach: (a version of) LRP
- Experiments
  - Getting Acquainted
  - Reference, Model and Random Prefixes
  - Exposure Bias and Source Contribution
  - Varying the Amount of Data
  - Training Stages

(A Bit of) the Training Process (work in progress)

# What is going to happen:

## The Trade-Off Between Source and Target

- Our Approach: (a version of) LRP
- Experiments
  - Getting Acquainted
  - Reference, Model and Random Prefixes
  - Exposure Bias and Source Contribution
  - Varying the Amount of Data
  - Training Stages

(A Bit of) the Training Process (work in progress)



# Model Prefixes are Simpler

Compared to references,  
beam search translations:

- contain fewer rare tokens

Burlot & Yvon, WMT 2018,  
Ott et al, ICML 2018

- have less reorderings

Burlot & Yvon, WMT 2018,  
Zhou et al, ICLR 2020

- are simpler syntactically

Burlot & Yvon, WMT 2018

# Model Prefixes are Simpler

Compared to references,  
beam search translations:

- contain fewer rare tokens

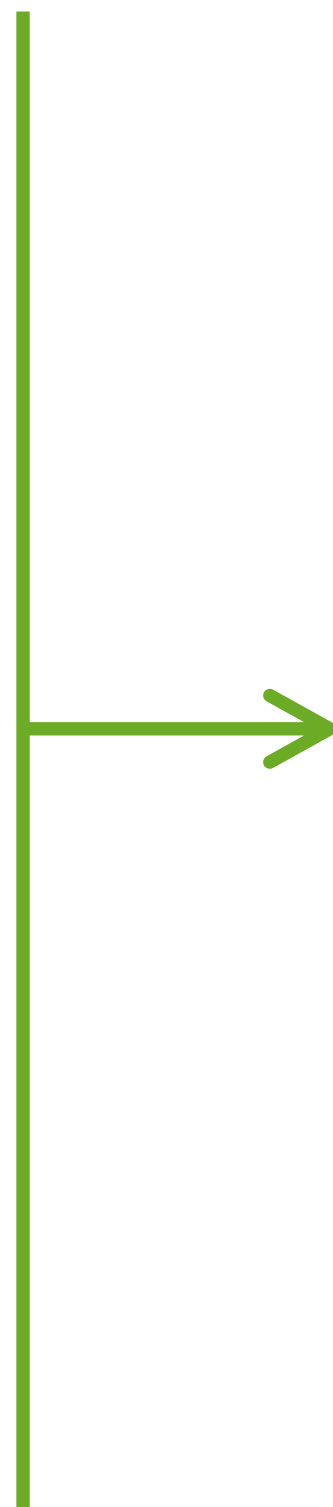
Burlot & Yvon, WMT 2018,  
Ott et al, ICML 2018

- have less reorderings

Burlot & Yvon, WMT 2018,  
Zhou et al, ICLR 2020

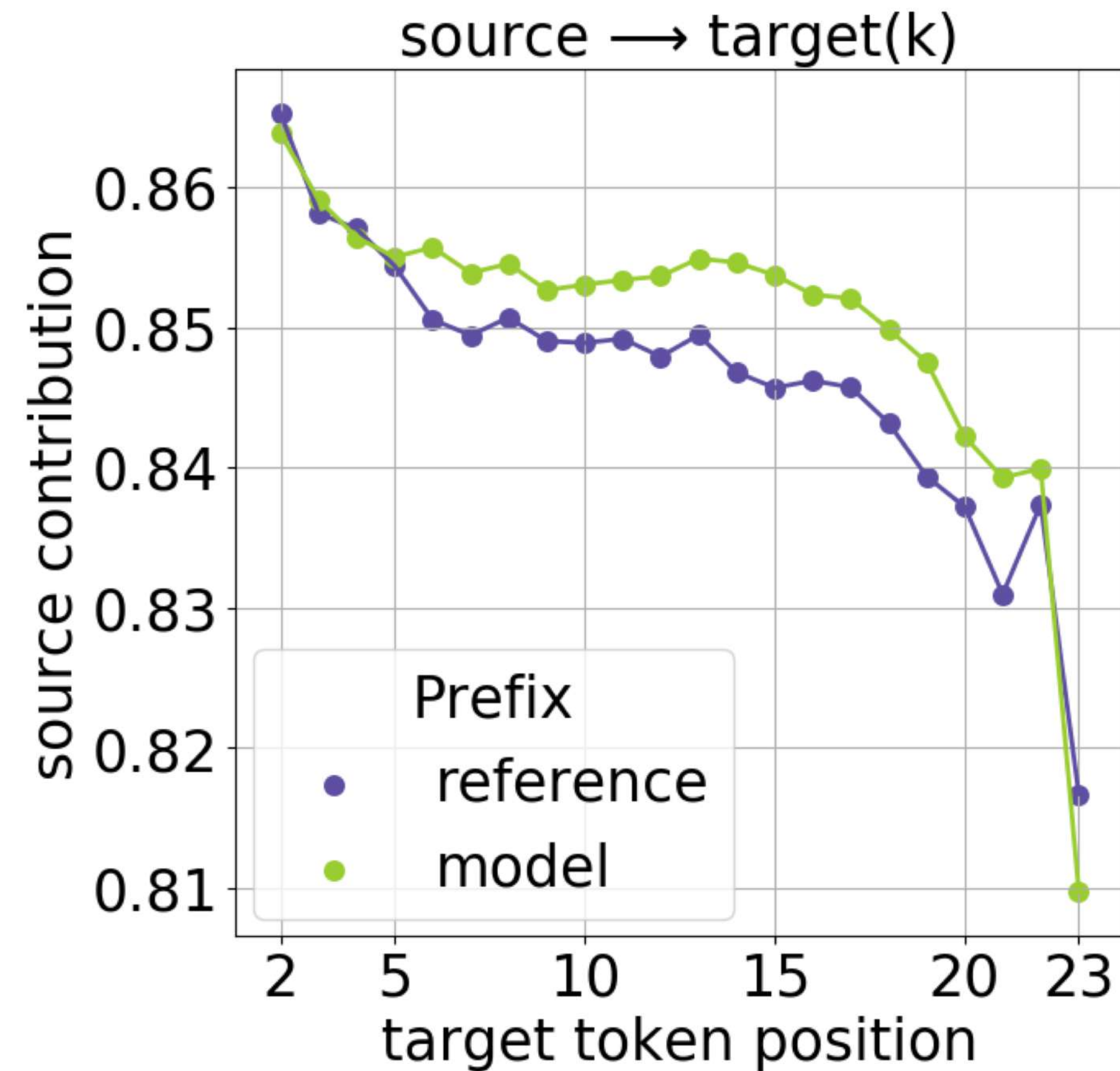
- are simpler syntactically

Burlot & Yvon, WMT 2018



Model-generated  
translations are **simpler**  
than references

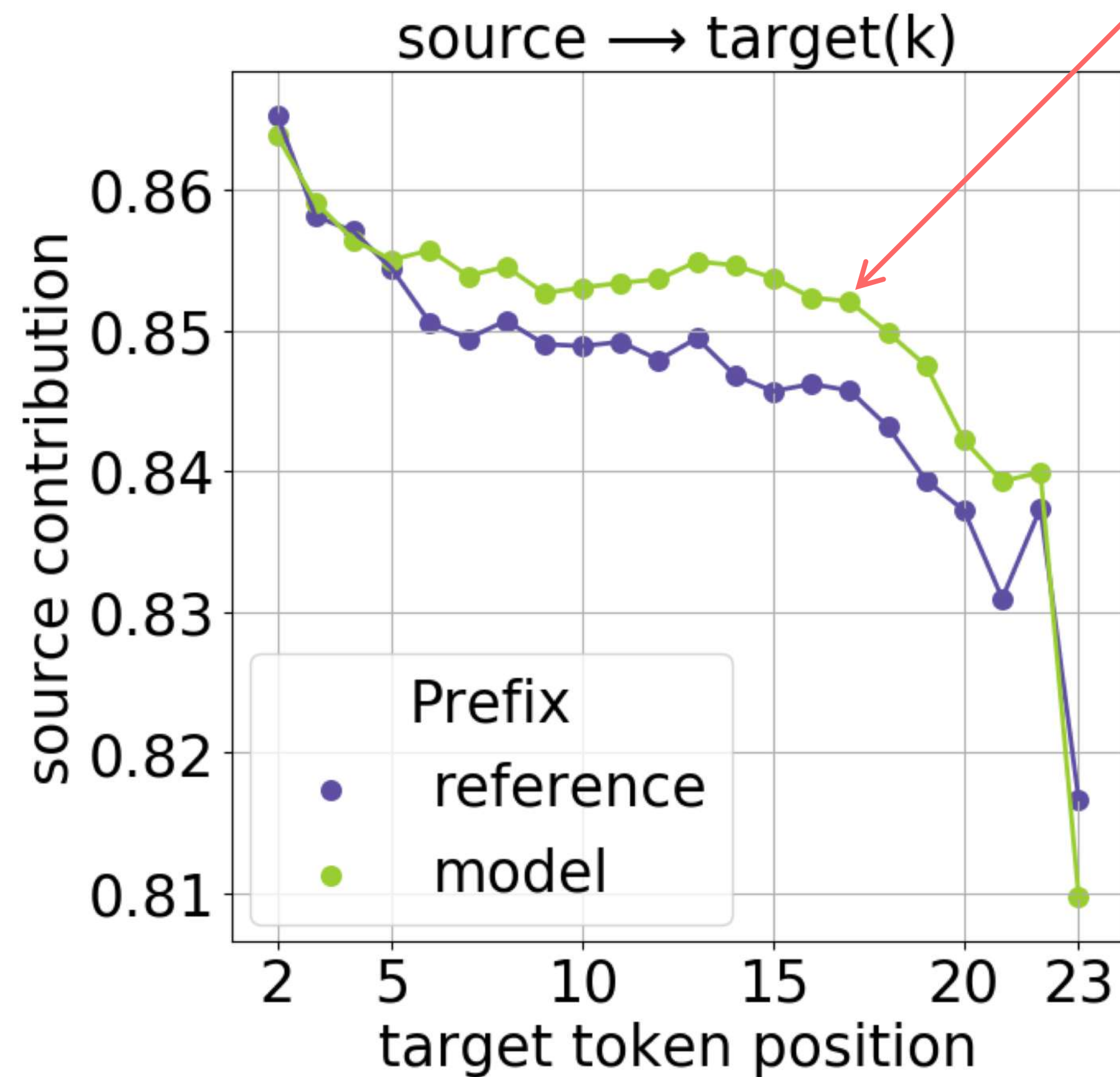
# Model vs Reference Prefixes



# Model vs Reference Prefixes

With model-generated prefixes:

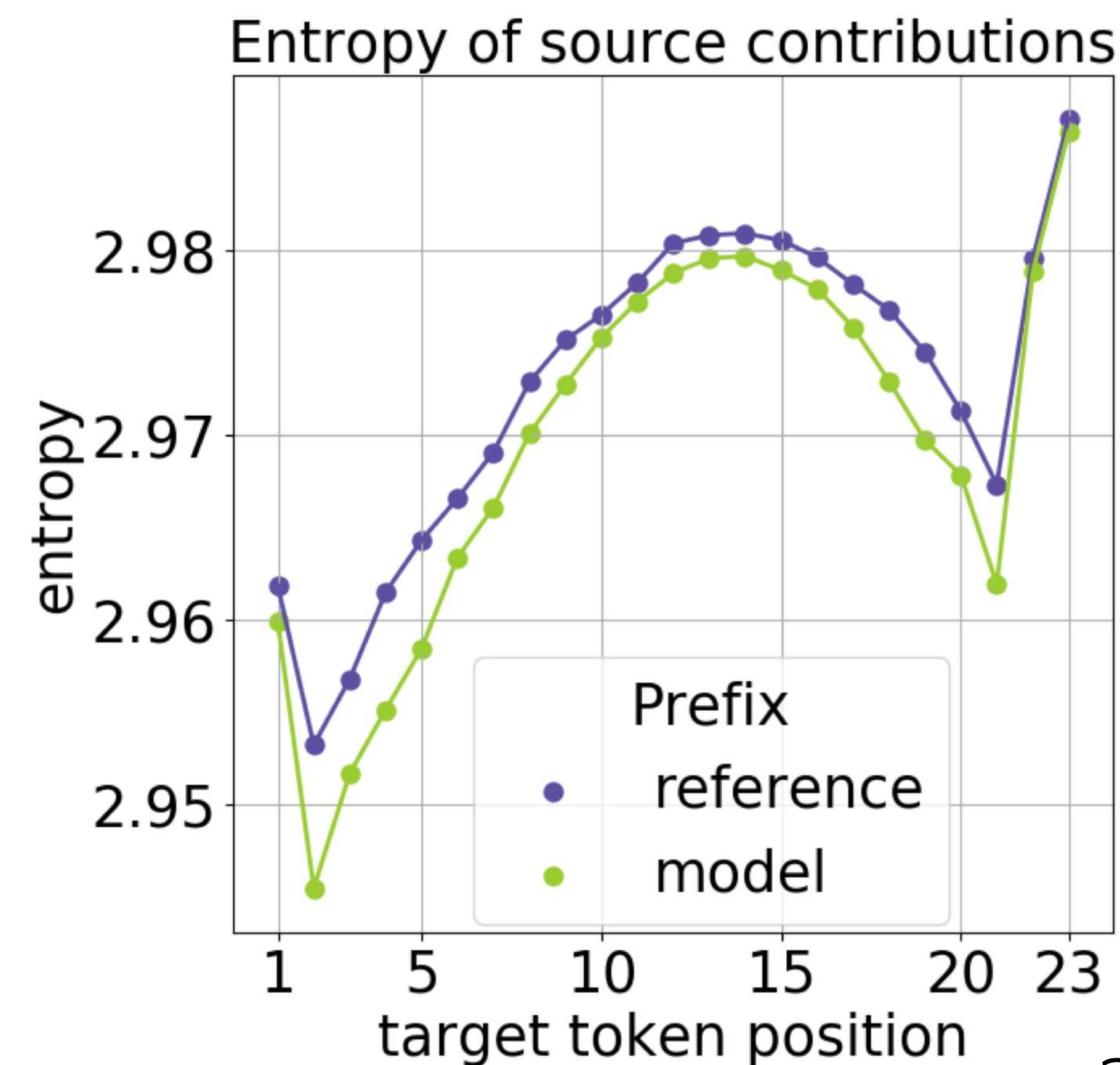
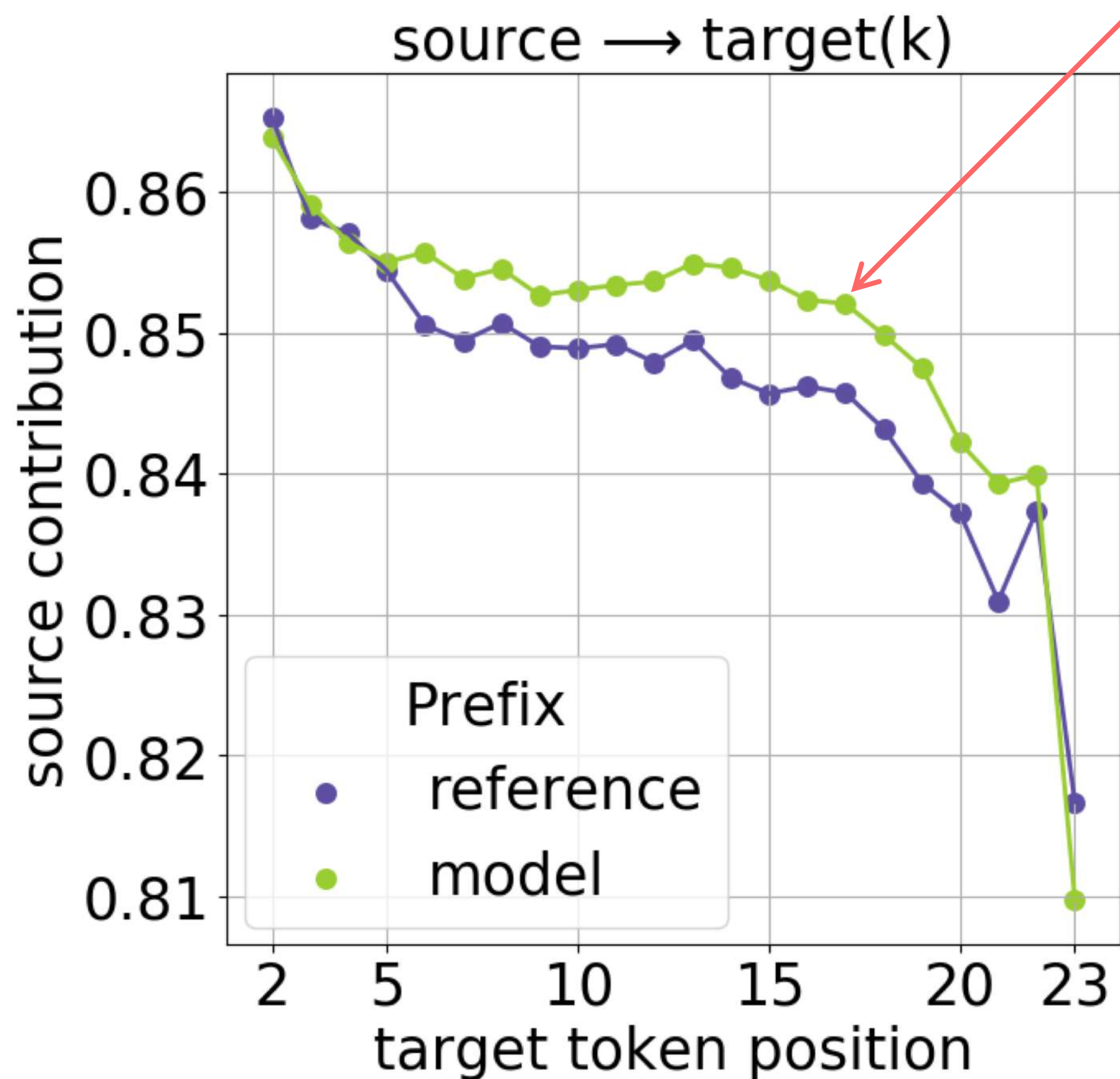
- source is used more



# Model vs Reference Prefixes

With model-generated prefixes:

- source is used more

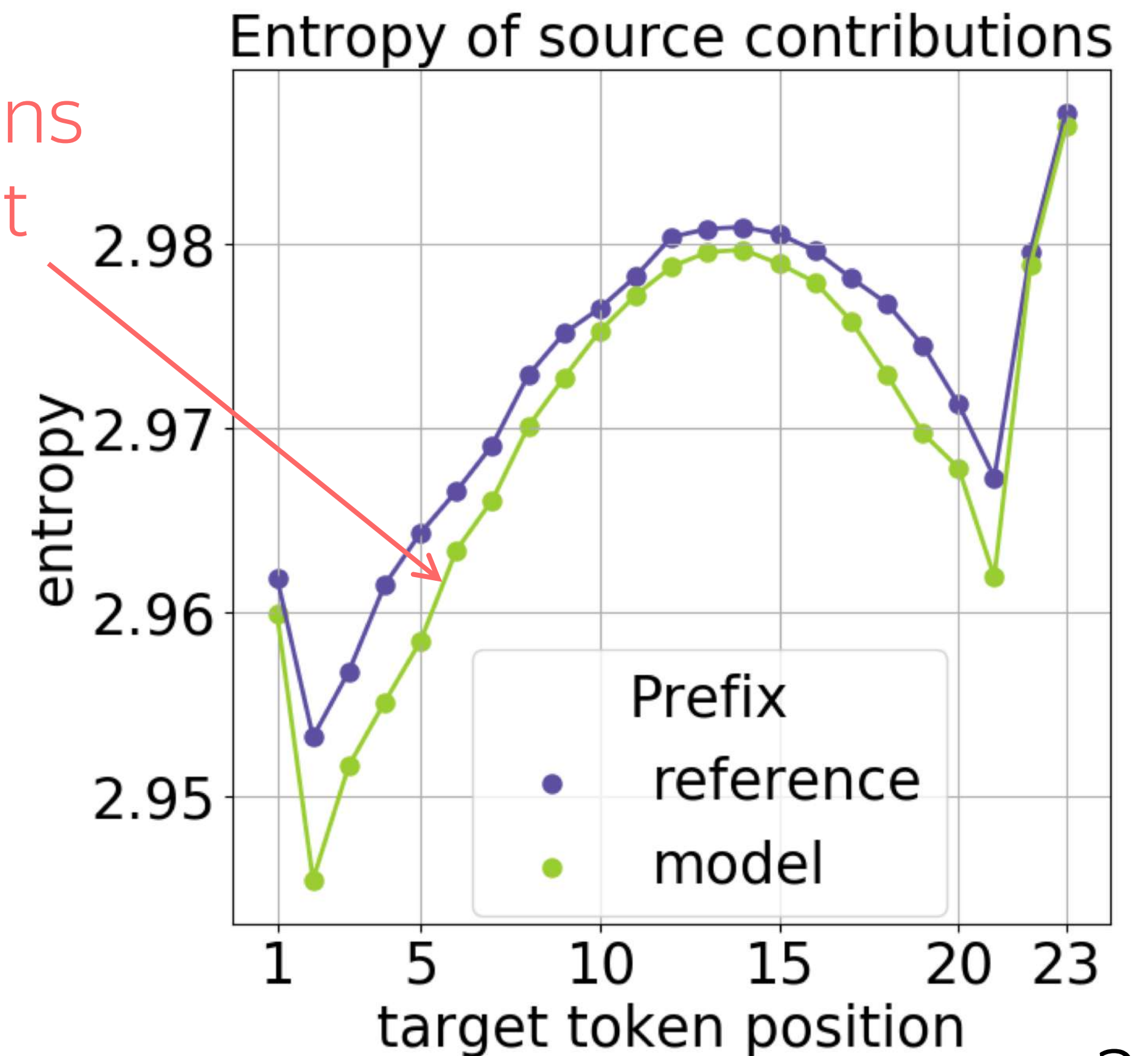
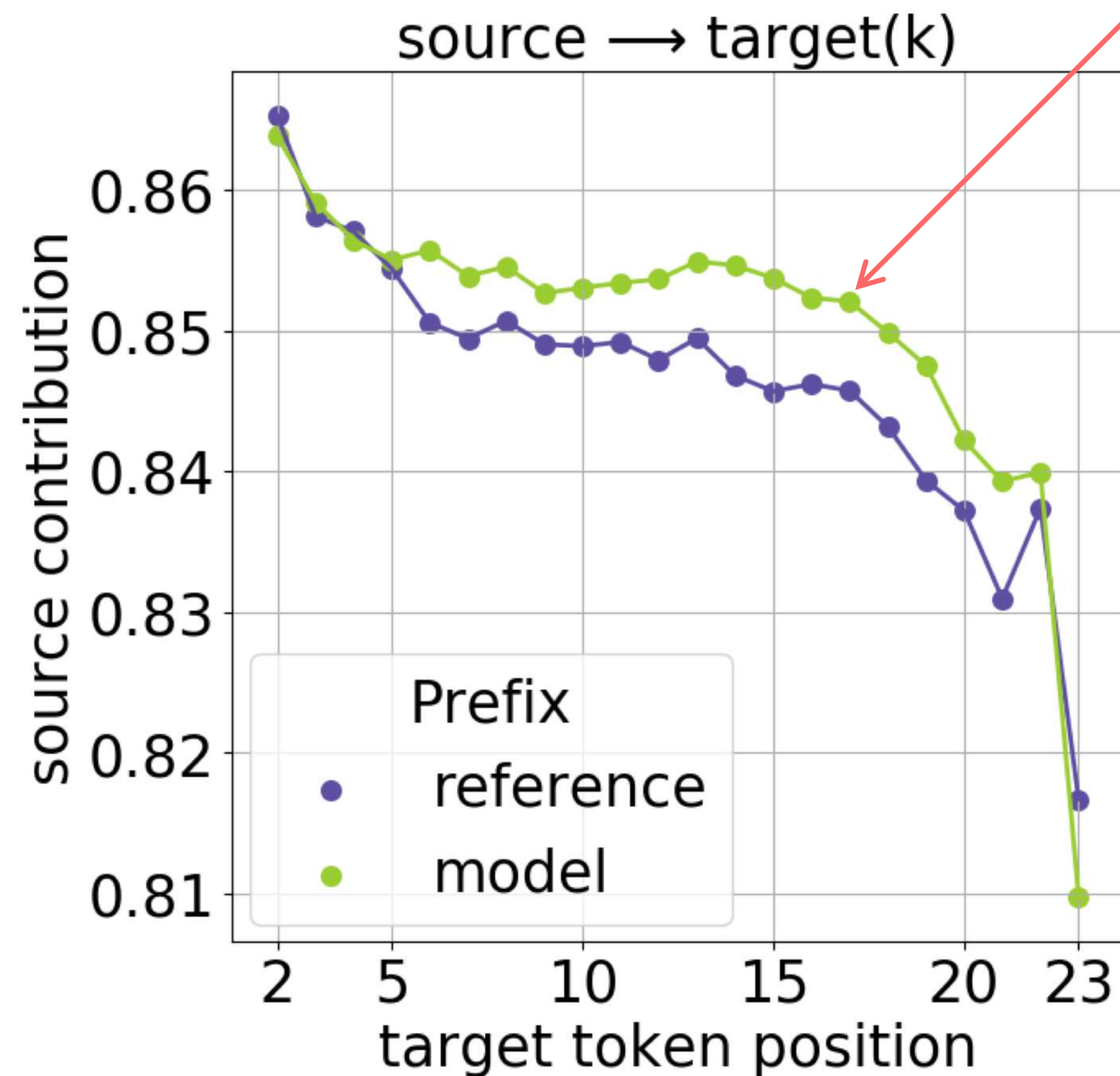




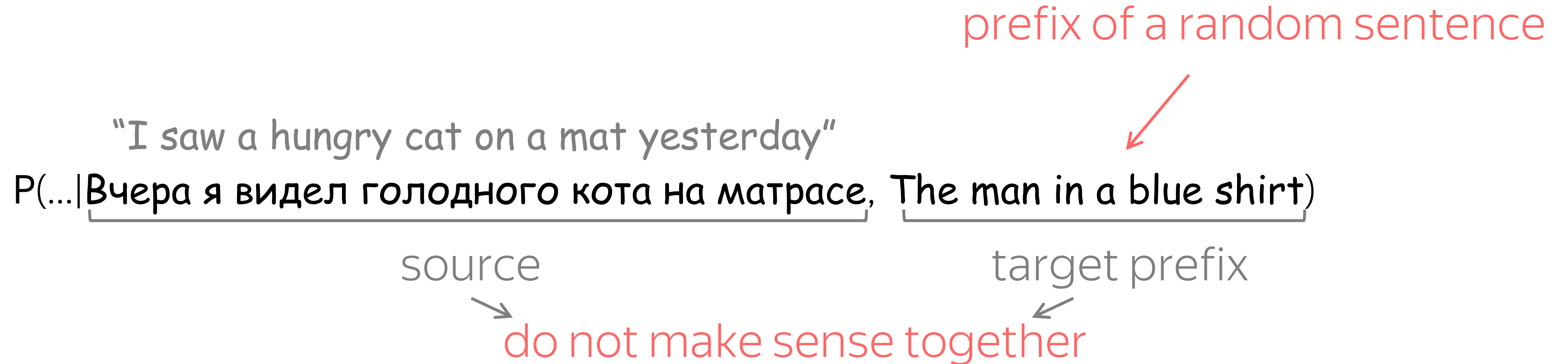
# Model vs Reference Prefixes

With model-generated prefixes:

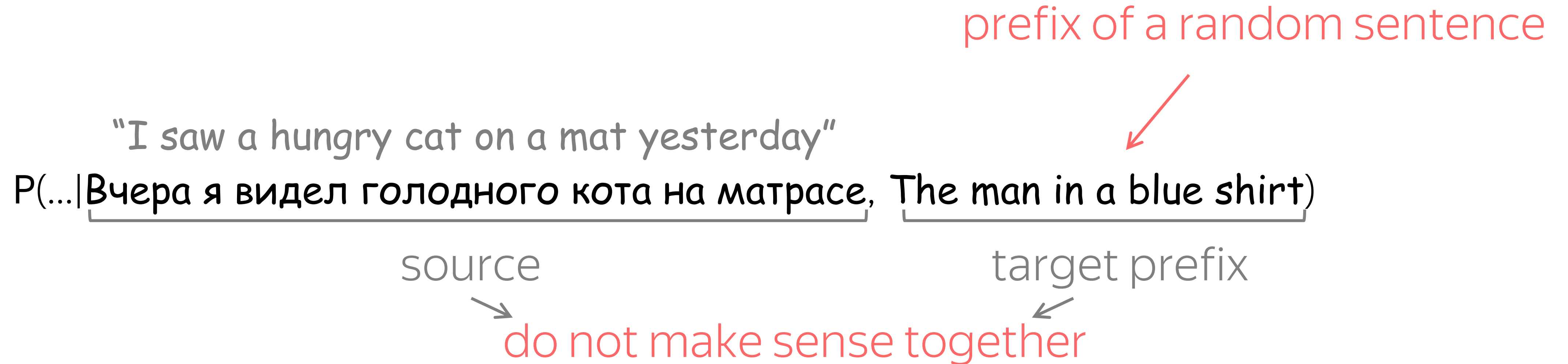
- source is used more
- source contributions are more confident



# Random vs Reference Prefixes



# Random vs Reference Prefixes



Why random prefixes?

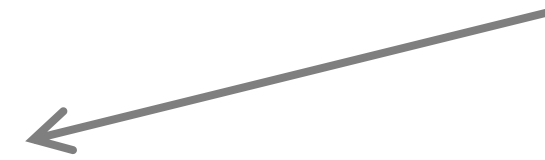
- We want to understand what happens when a model is hallucinating
- Random prefixes is a simple way to simulate hallucination mode

# Random vs Reference Prefixes

Previous work

# Random vs Reference Prefixes

Previous work

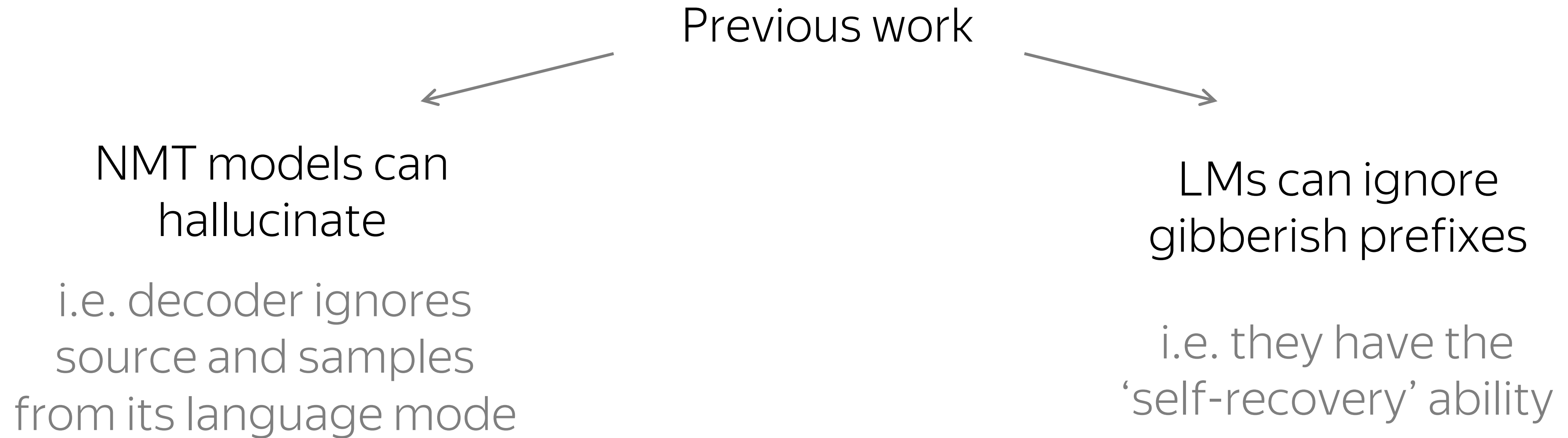


NMT models can  
hallucinate

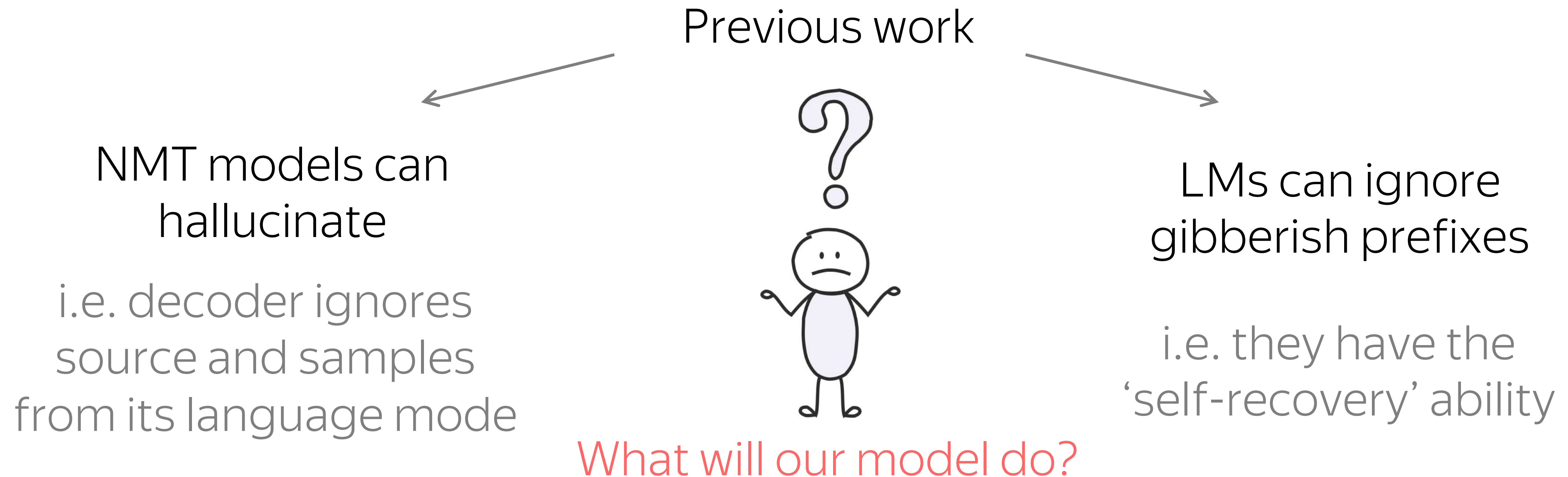
i.e. decoder ignores  
source and samples  
from its language mode



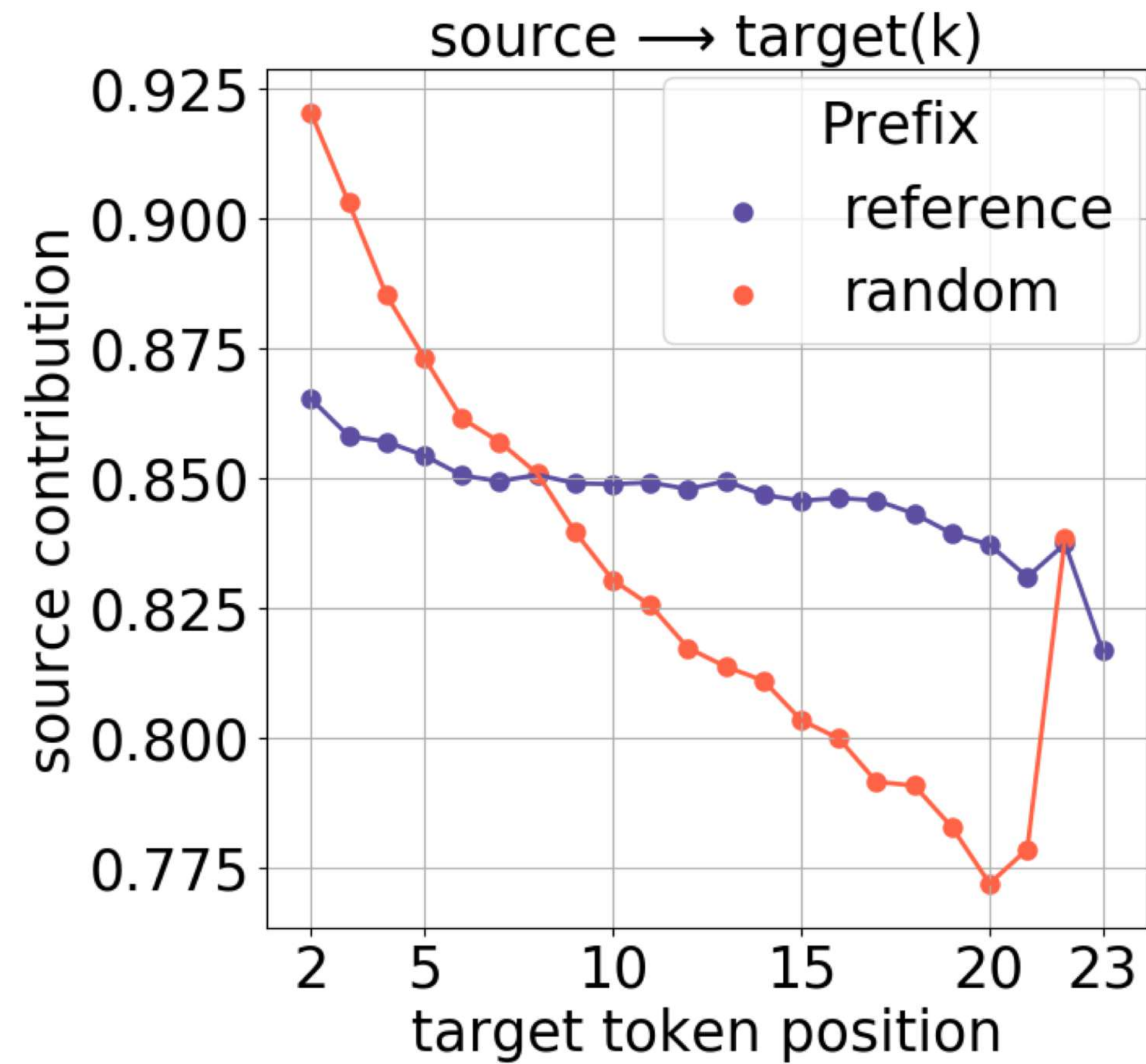
# Random vs Reference Prefixes



# Random vs Reference Prefixes

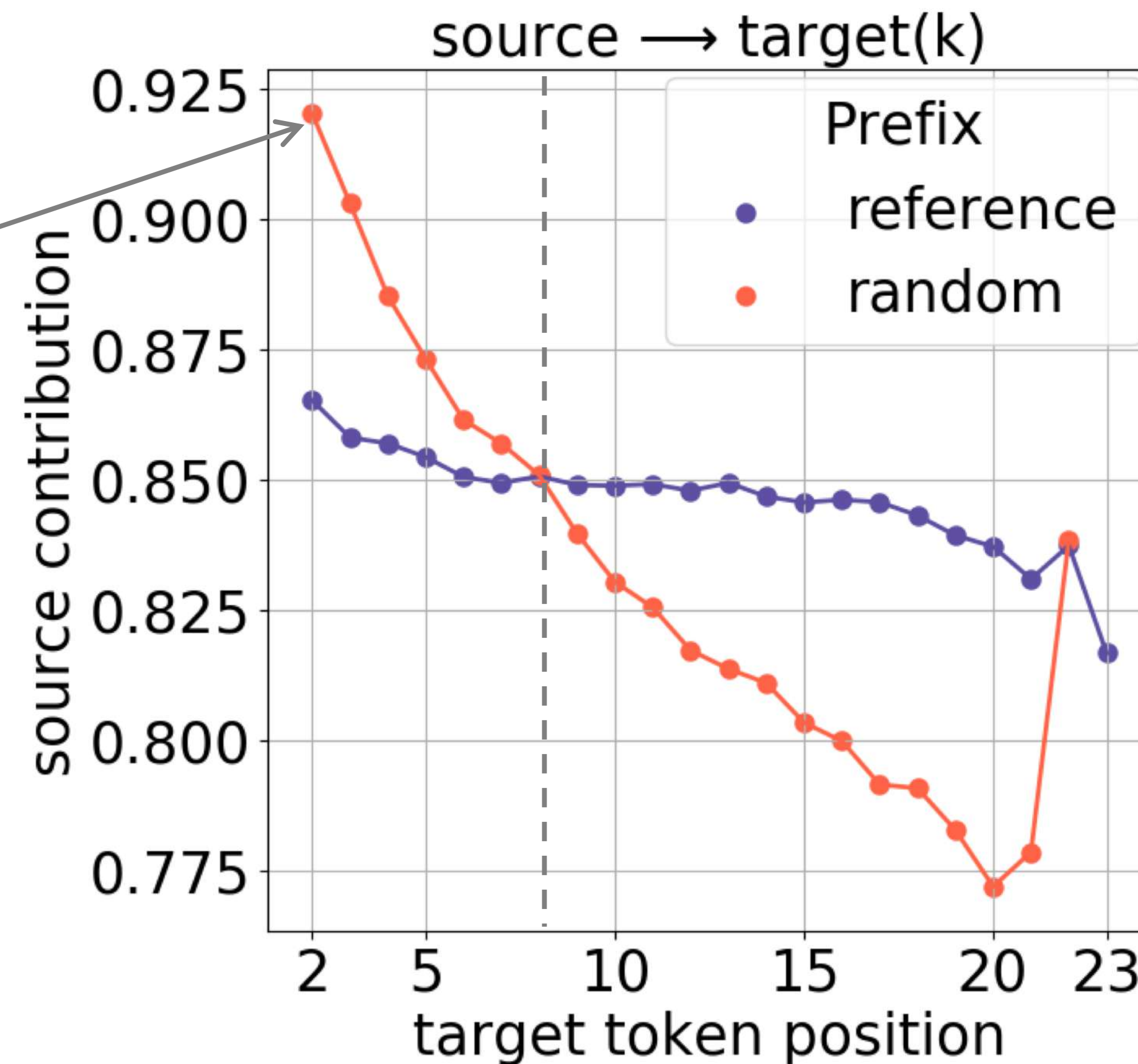


# Random vs Reference Prefixes



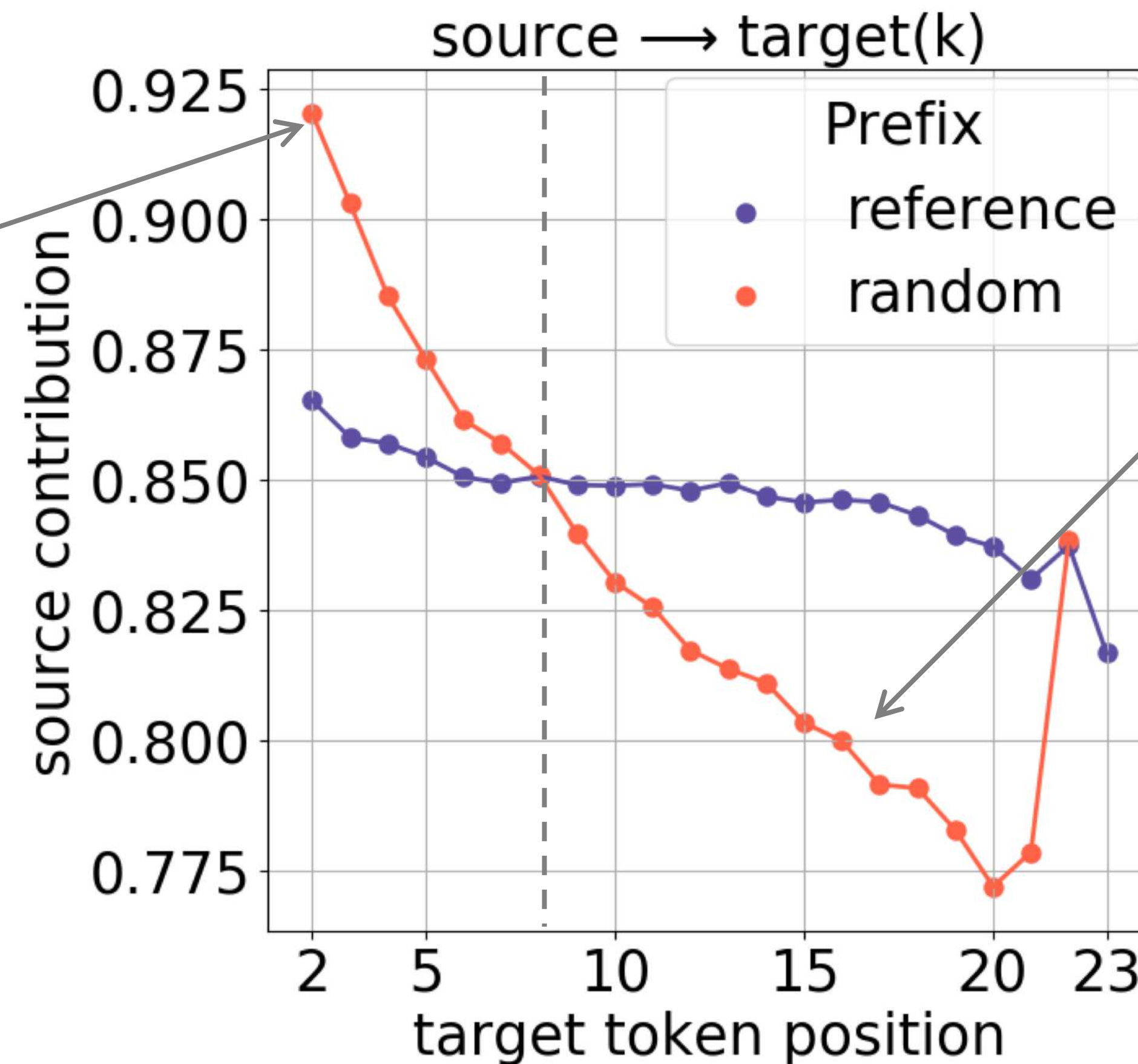
# Random vs Reference Prefixes

When a random prefix is short, the model “recovers”: it ignores the prefix (very high source contribution)



# Random vs Reference Prefixes

When a random prefix is short, the model “recovers”: it ignores the prefix  
(very high source contribution)



When a random prefix is long, the model “hallucinates”: it ignores the source  
(very low source contribution)



# Summary: Different Prefixes

## Reference vs Model prefixes

- a model uses source more and does it more confidently
- probably because model-generated prefixes are simpler

## Reference vs Random prefixes

- if a random prefix is short, a model ignores the prefix
- if a random prefix is long, a model ignores the source

# What is going to happen:

## The Trade-Off Between Source and Target

- Our Approach: (a version of) LRP
- Experiments
  - Getting Acquainted
  - Reference, Model and Random Prefixes
  - Exposure Bias and Source Contribution
  - Varying the Amount of Data
  - Training Stages

(A Bit of) the Training Process (work in progress)

# What is going to happen:

## The Trade-Off Between Source and Target

- Our Approach: (a version of) LRP
- Experiments
  - Getting Acquainted
  - Reference, Model and Random Prefixes
  - Exposure Bias and Source Contribution
  - Varying the Amount of Data
  - Training Stages

(A Bit of) the Training Process (work in progress)

Experiments:

Exposure Bias and Source Contributions



# Exposure bias and Hallucinations

Training objective	Exposure bias
MLE (standard)	suffer
Minimum Risk Training	do not suffer



# Exposure bias and Hallucinations

Training objective	Exposure bias	Hallucinations
MLE (standard)	suffer	suffer
Minimum Risk Training	do not suffer	suffer less

# Exposure bias and Hallucinations

Training objective	Exposure bias	Hallucinations	Hypothesis:
MLE (standard)	suffer	suffer	Exposure bias leads to over-reliance on target history
Minimum Risk Training	do not suffer	suffer less	

# Exposure bias and Hallucinations

Previous work ([Wang & Sennrich, ACL 2020](#))

Training objective	Exposure bias	Hallucinations	Hypothesis:
MLE (standard)	suffer	suffer	Exposure bias leads to over-reliance on target history
Minimum Risk Training	do not suffer	suffer less	

# Exposure bias and Hallucinations

Previous work (Wang & Sennrich, ACL 2020)

We

Training objective	Exposure bias	Hallucinations	Hypothesis:	
MLE (standard)	suffer	suffer	Exposure bias leads to over-reliance on target history	→ Let's measure!
Minimum Risk Training	do not suffer	suffer less		

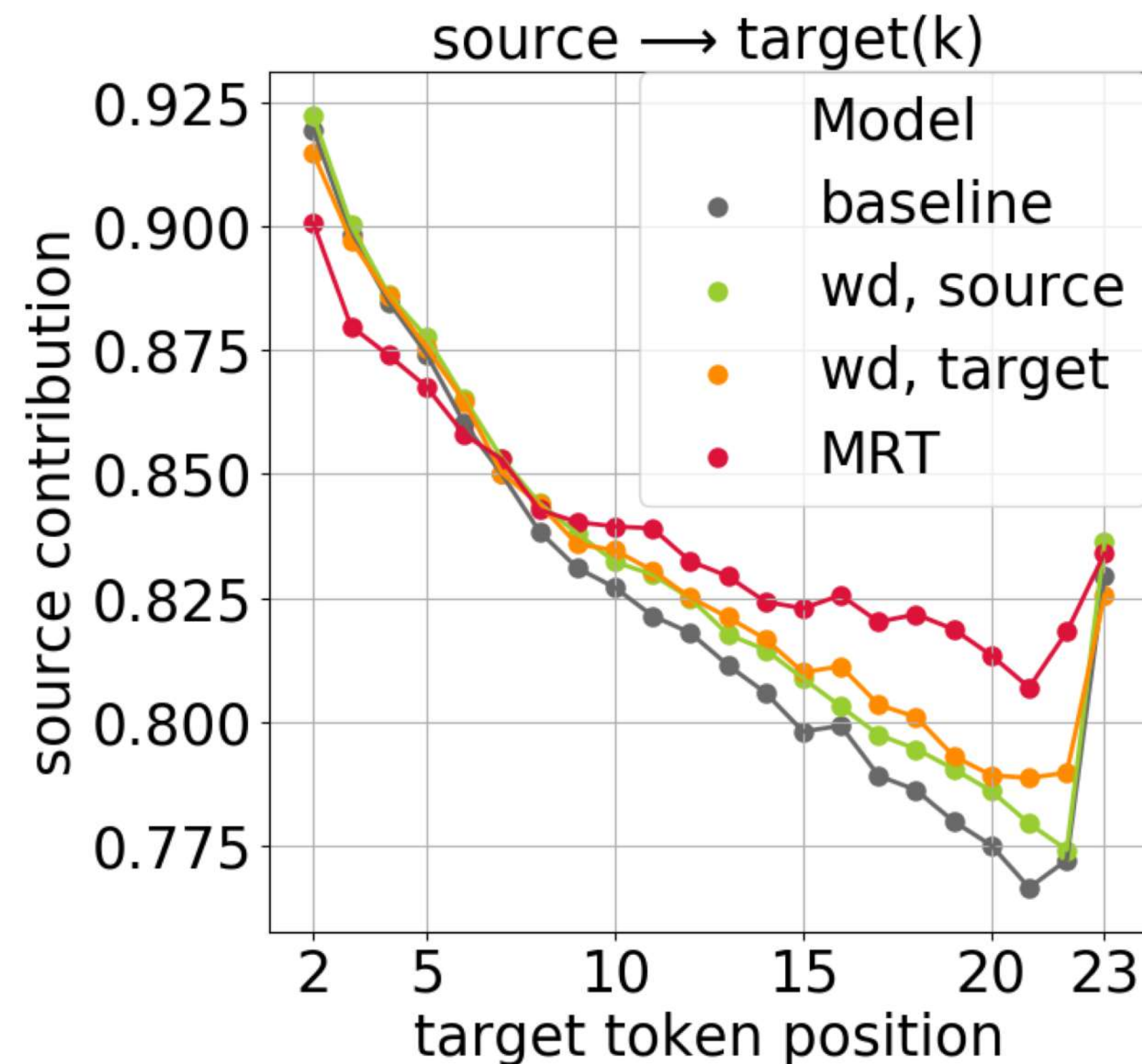
# Random prefix, source contribution

Model	Exposure bias
baseline	suffer
word dropout, source	suffer
word dropout, target	suffer a bit less
Minimum Risk Training	do not suffer



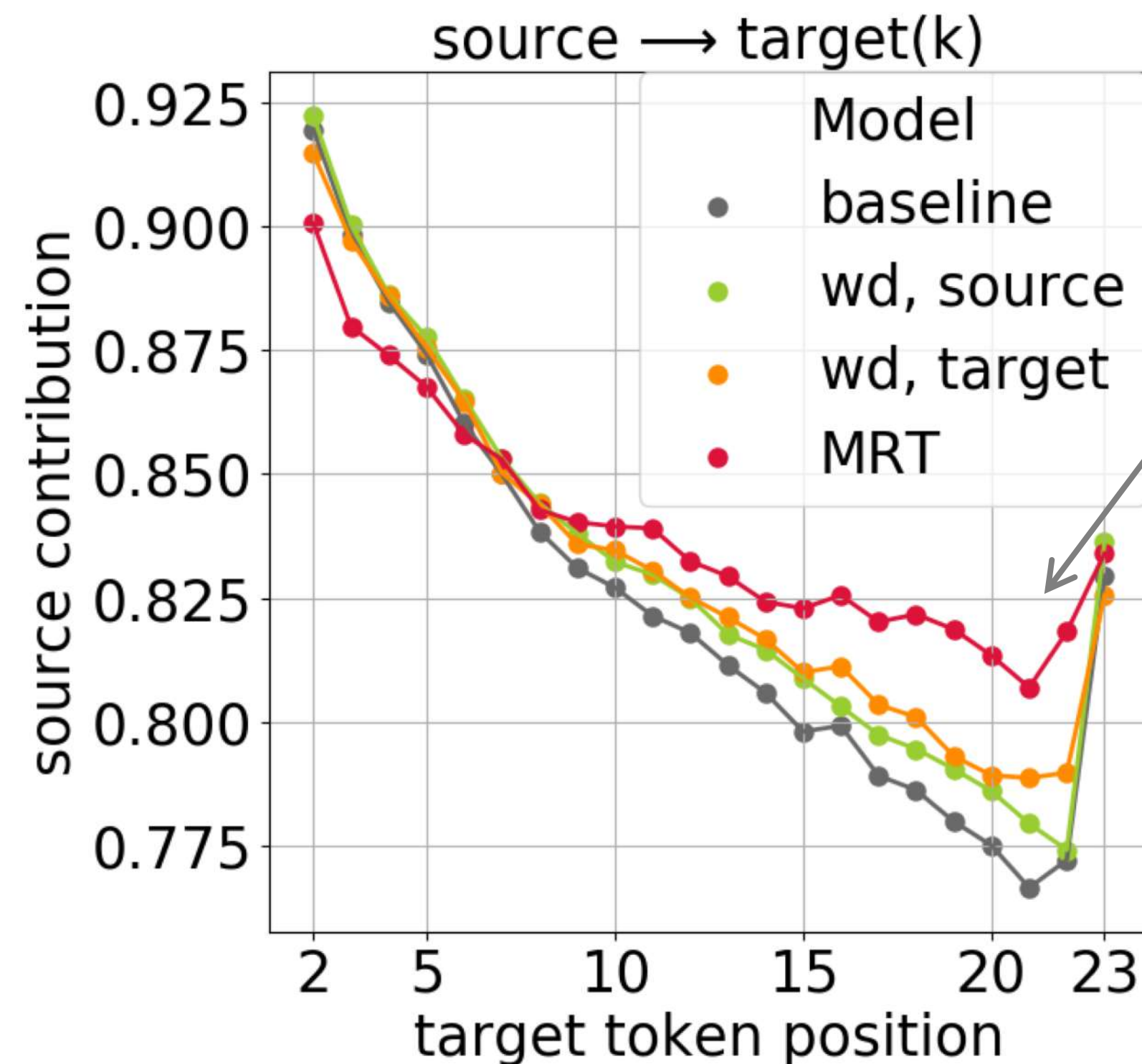
# Random prefix, source contribution

Model	Exposure bias
baseline	suffer
word dropout, source	suffer
word dropout, target	suffer a bit less
Minimum Risk Training	do not suffer



# Random prefix, source contribution

Model	Exposure bias
baseline	suffer
word dropout, source	suffer
word dropout, target	suffer a bit less
Minimum Risk Training	do not suffer

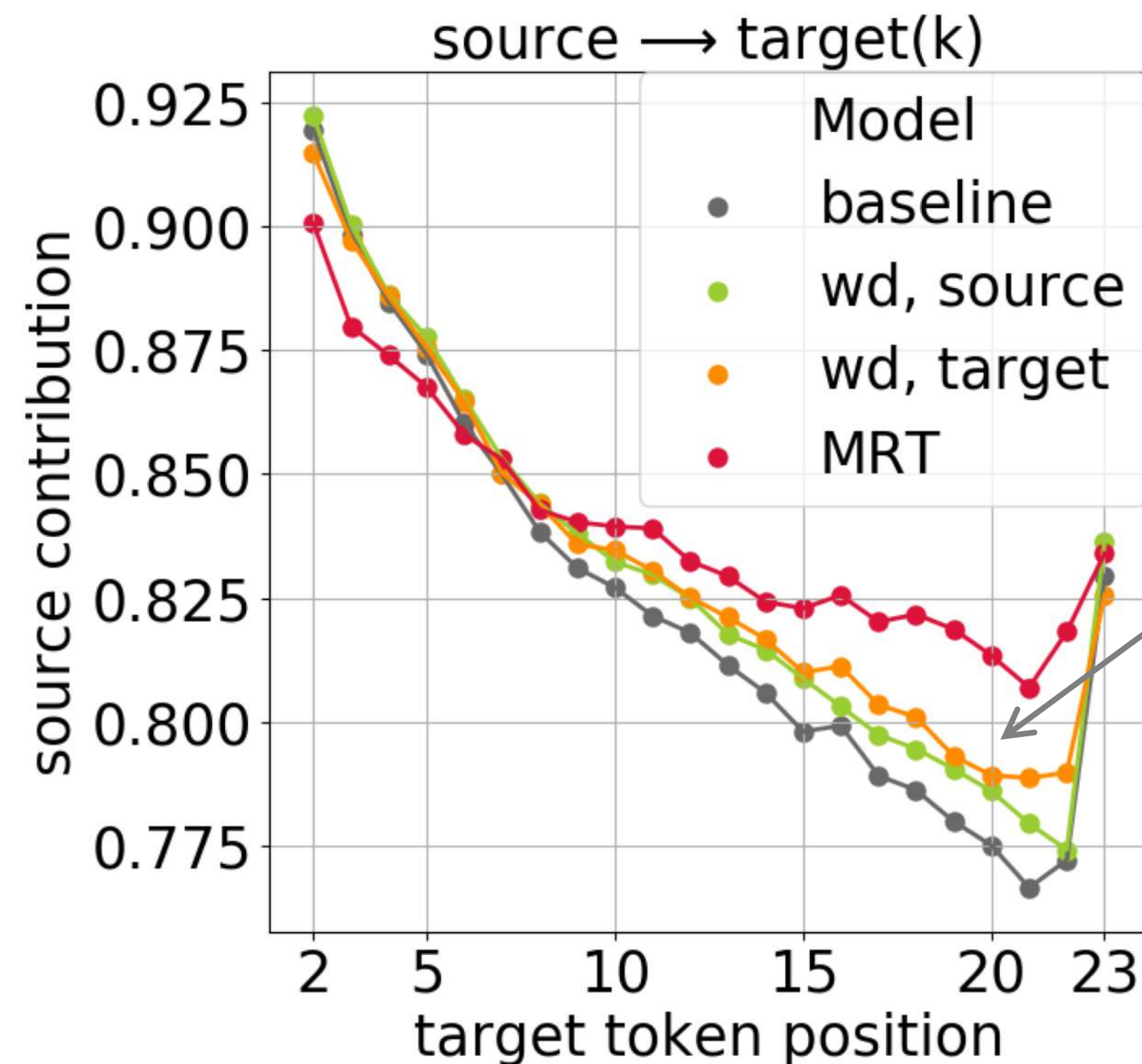


With MRT, models ignore the source less than any other model

# Random prefix, source contribution

With word dropout, models:

Model	Exposure bias
baseline	suffer
word dropout, source	suffer
word dropout, target	suffer a bit less
Minimum Risk Training	do not suffer

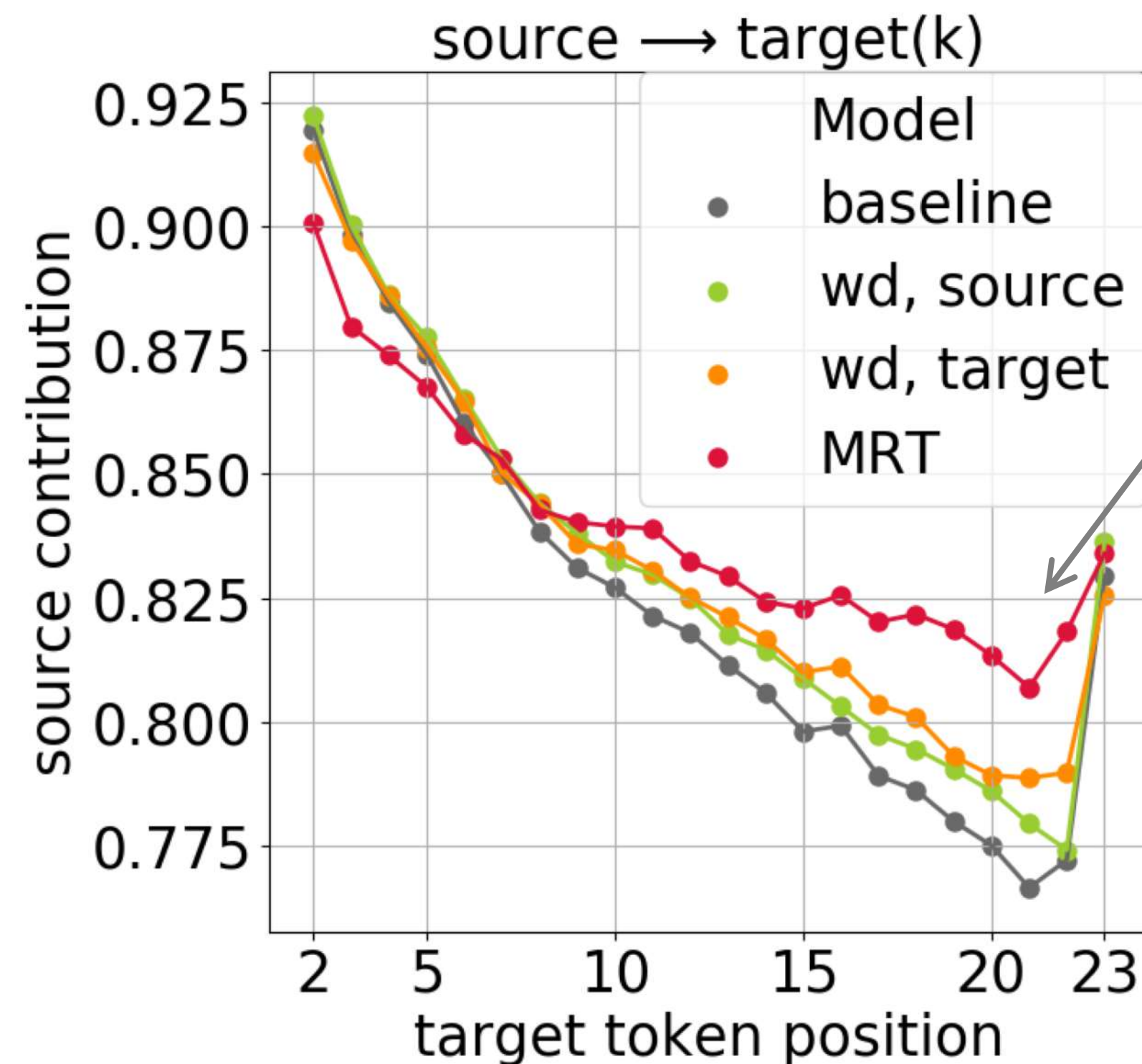


- ignore the source a bit less than the baseline
- on the target side the effect is larger



# Random prefix, source contribution

Model	Exposure bias
baseline	suffer
word dropout, source	suffer
word dropout, target	suffer a bit less
Minimum Risk Training	do not suffer

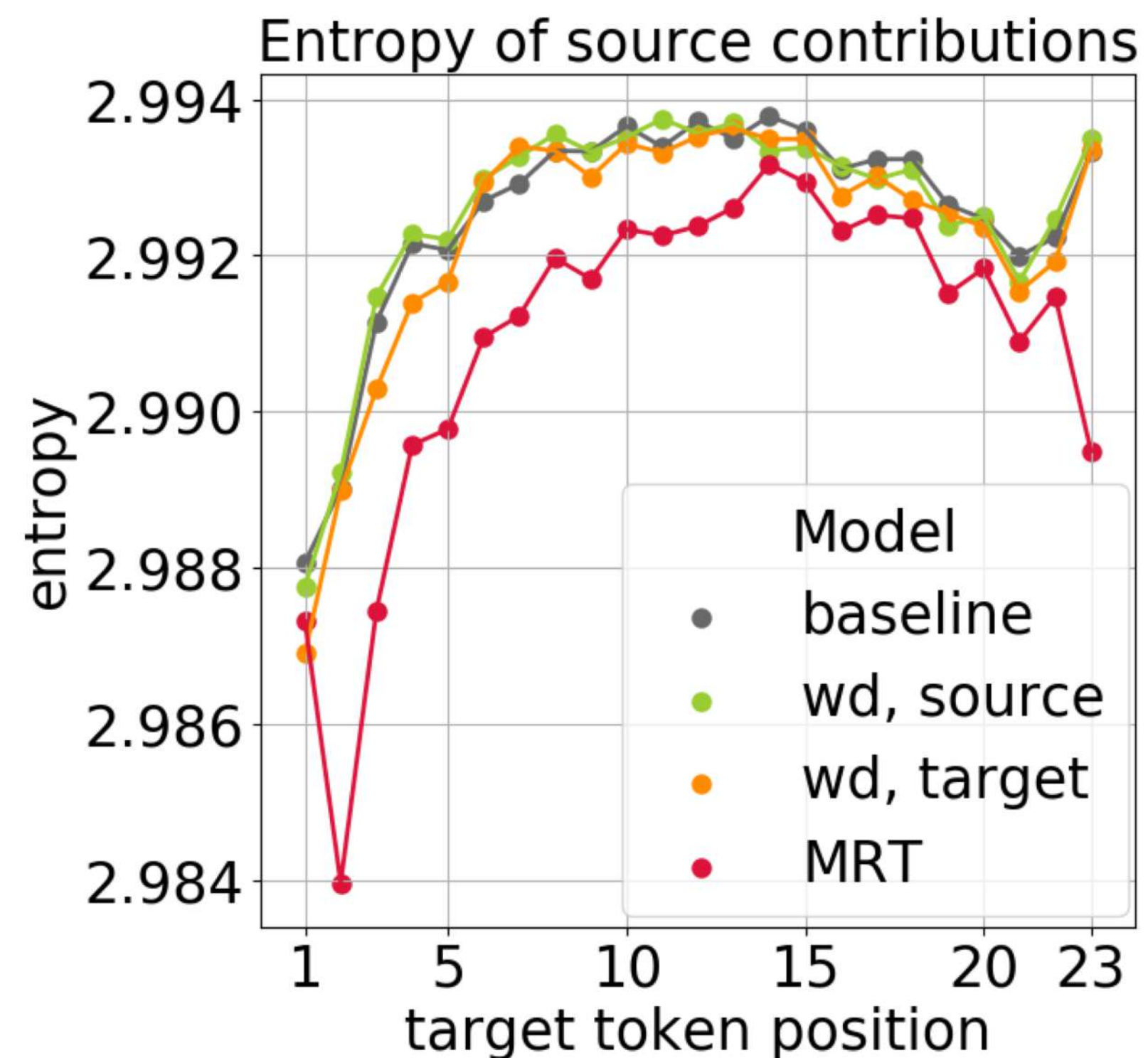


Models with alleviated exposure bias ignore the source less than other models

(when conditioned on random prefixes)

# Random prefix, entropy of source contributions

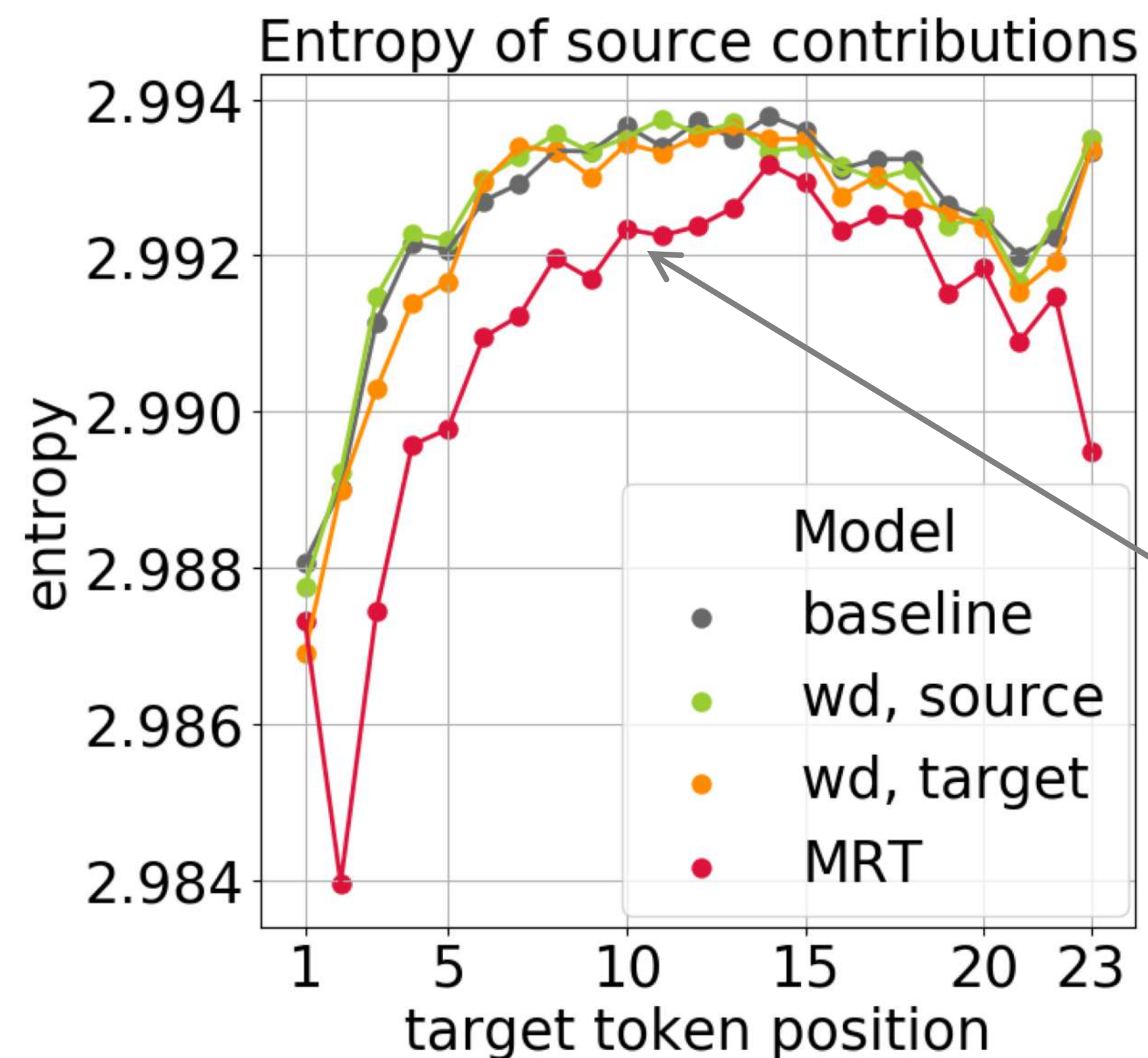
Model	Exposure bias
baseline	suffer
word dropout, source	suffer
word dropout, target	suffer a bit less
Minimum Risk Training	do not suffer





# Random prefix, entropy of source contributions

Model	Exposure bias
baseline	suffer
word dropout, source	suffer
word dropout, target	suffer a bit less
Minimum Risk Training	do not suffer



With MRT, models use source more confidently

# Summary: Exposure Bias and Hallucinations

Compared to models where the exposure bias is mitigated,

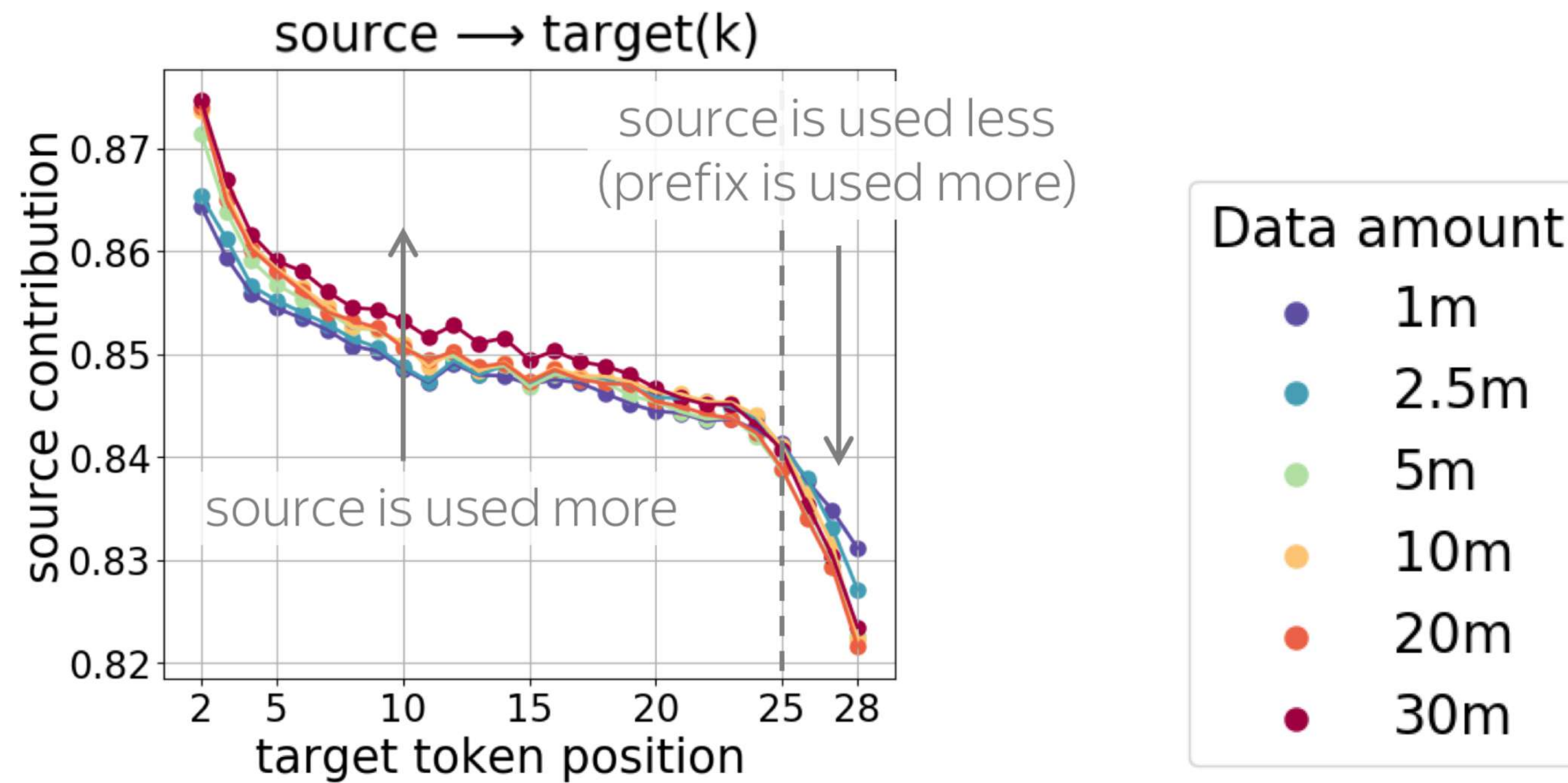
- Models suffering from exposure bias are more prone to over-relying on target history (and hence to hallucinating)

Experiments:

Varying the Amount of Data



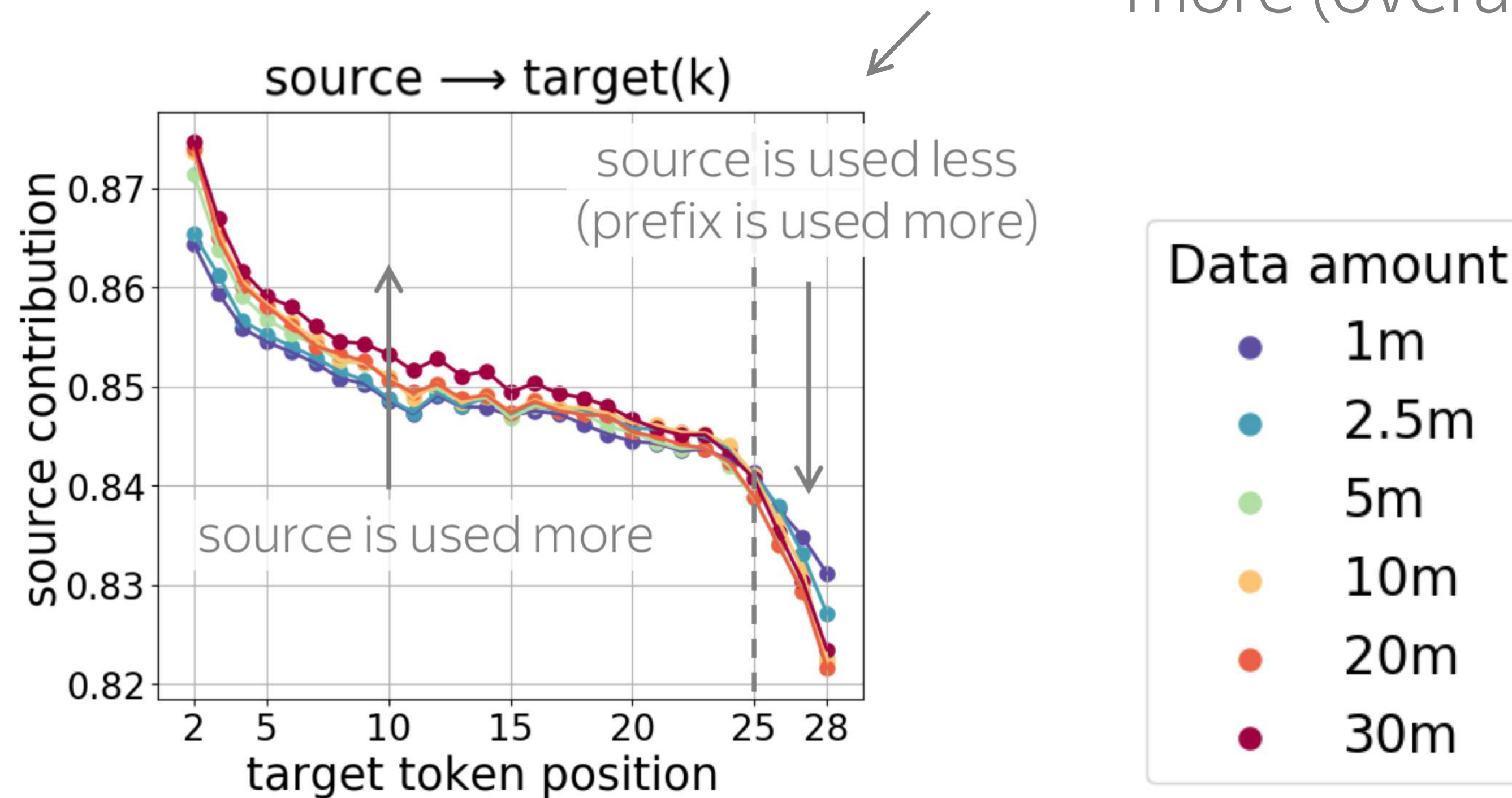
# Varying Amount of Training Data



# Varying Amount of Training Data

With more data, models use source:

- more (overall)

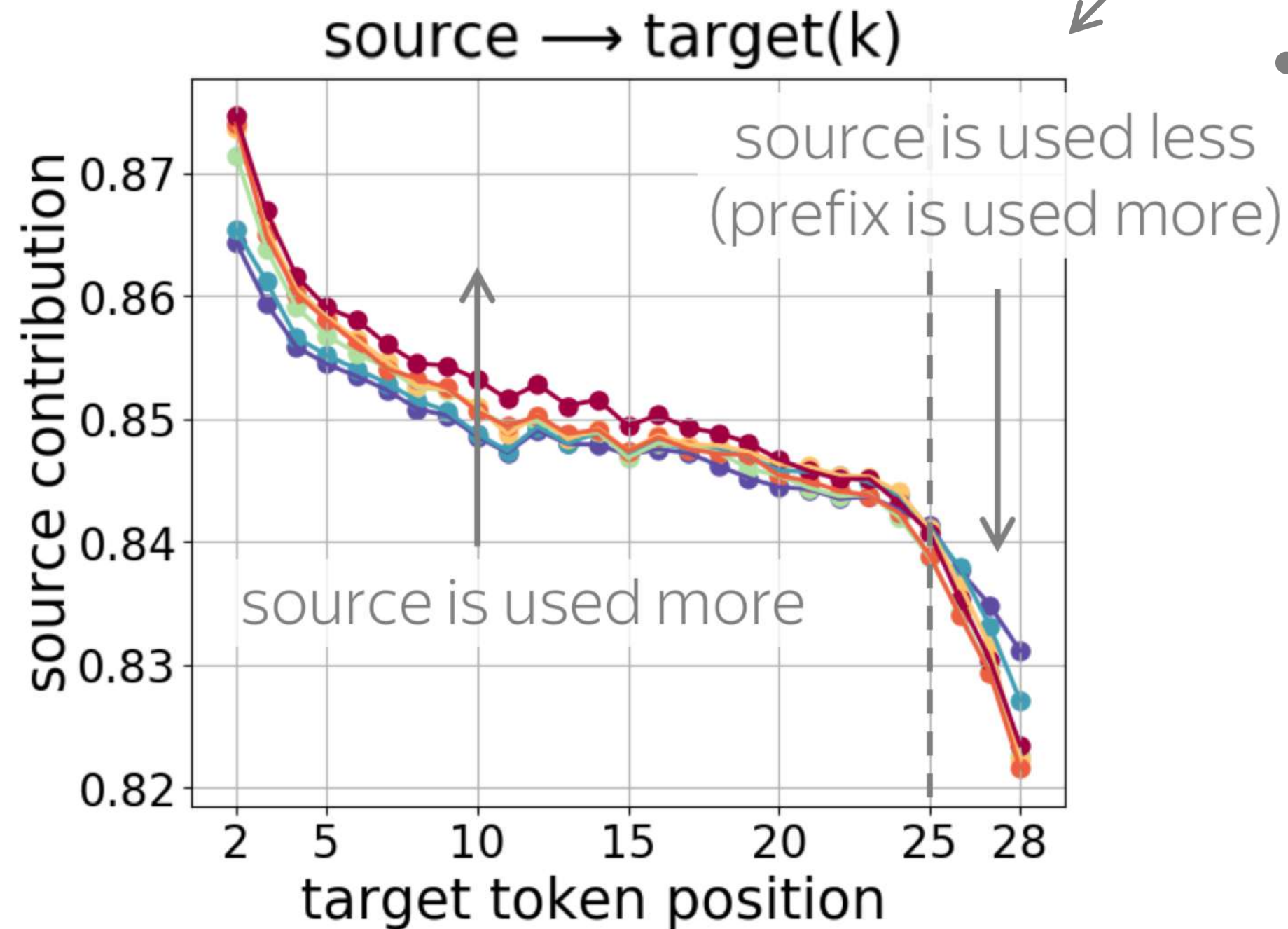




# Varying Amount of Training Data

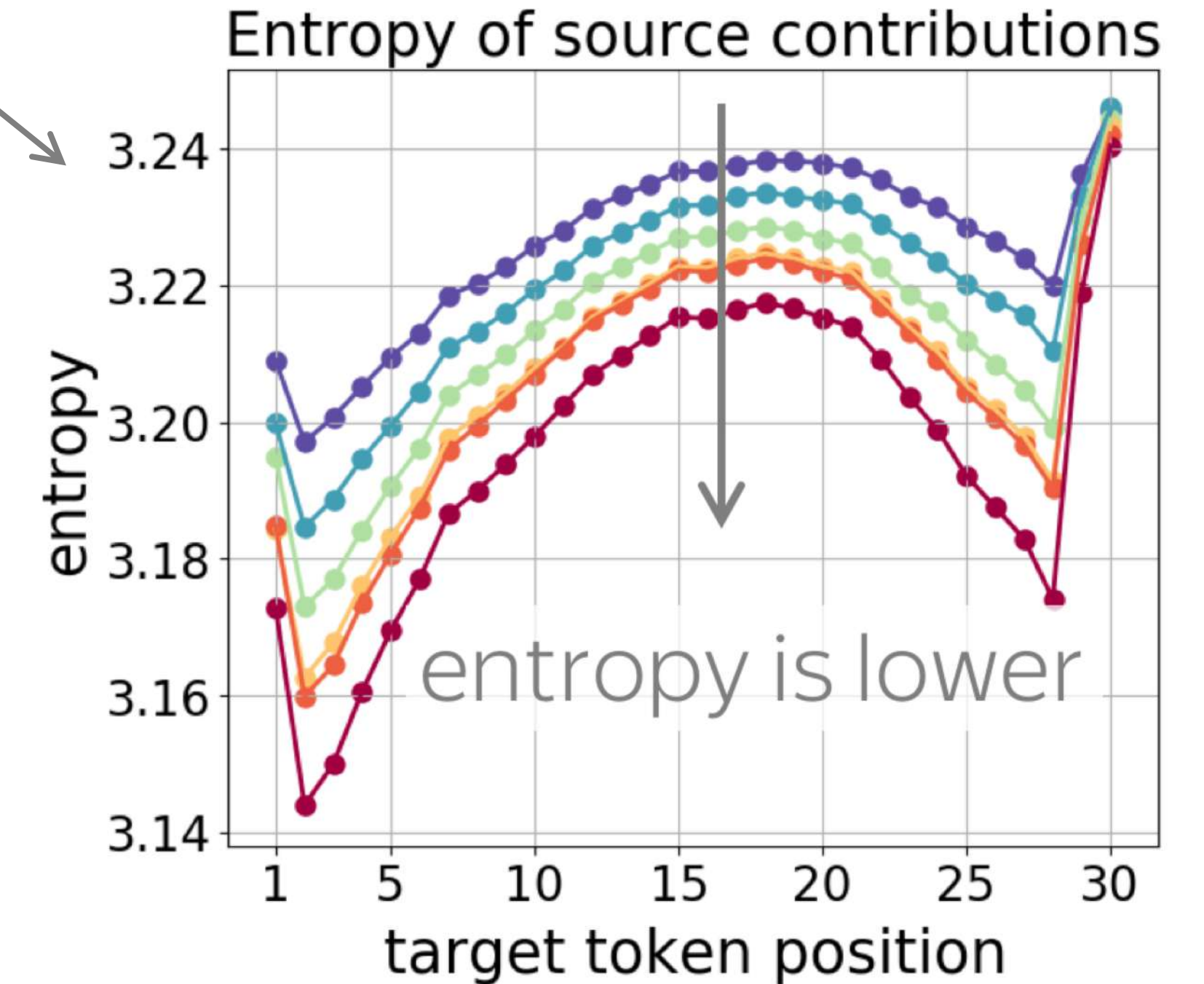
With more data, models use source:

- more (overall)
- more confidently



Data amount

- 1m
- 2.5m
- 5m
- 10m
- 20m
- 30m



# What is going to happen:

## The Trade-Off Between Source and Target

- Our Approach: (a version of) LRP
- Experiments
  - Getting Acquainted
  - Reference, Model and Random Prefixes
  - Exposure Bias and Source Contribution
  - Varying the Amount of Data
  - Training Stages

(A Bit of) the Training Process (work in progress)

# What is going to happen:

## The Trade-Off Between Source and Target

- Our Approach: (a version of) LRP
- Experiments
  - Getting Acquainted
  - Reference, Model and Random Prefixes
  - Exposure Bias and Source Contribution
  - Varying the Amount of Data
  - Training Stages

(A Bit of) the Training Process (work in progress)

# Experiments: Training Stages



# The Training Timeline

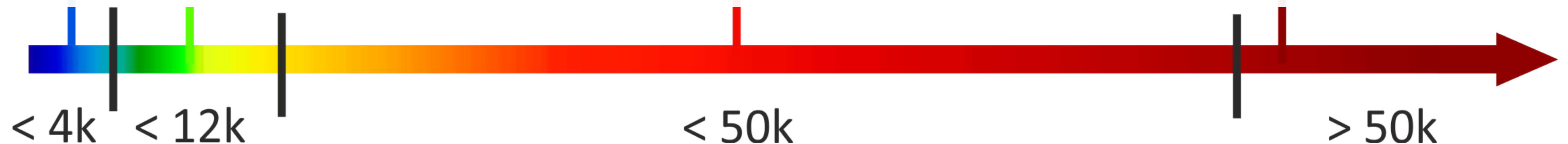
- a lot of change
- more change for early positions
- source influence decreases
- entropy of contributions decreases

- contributions converge, small changes
- equal change across positions
- source influence increases
- entropy of contributions increases

We will be uncovering this experiment by experiment

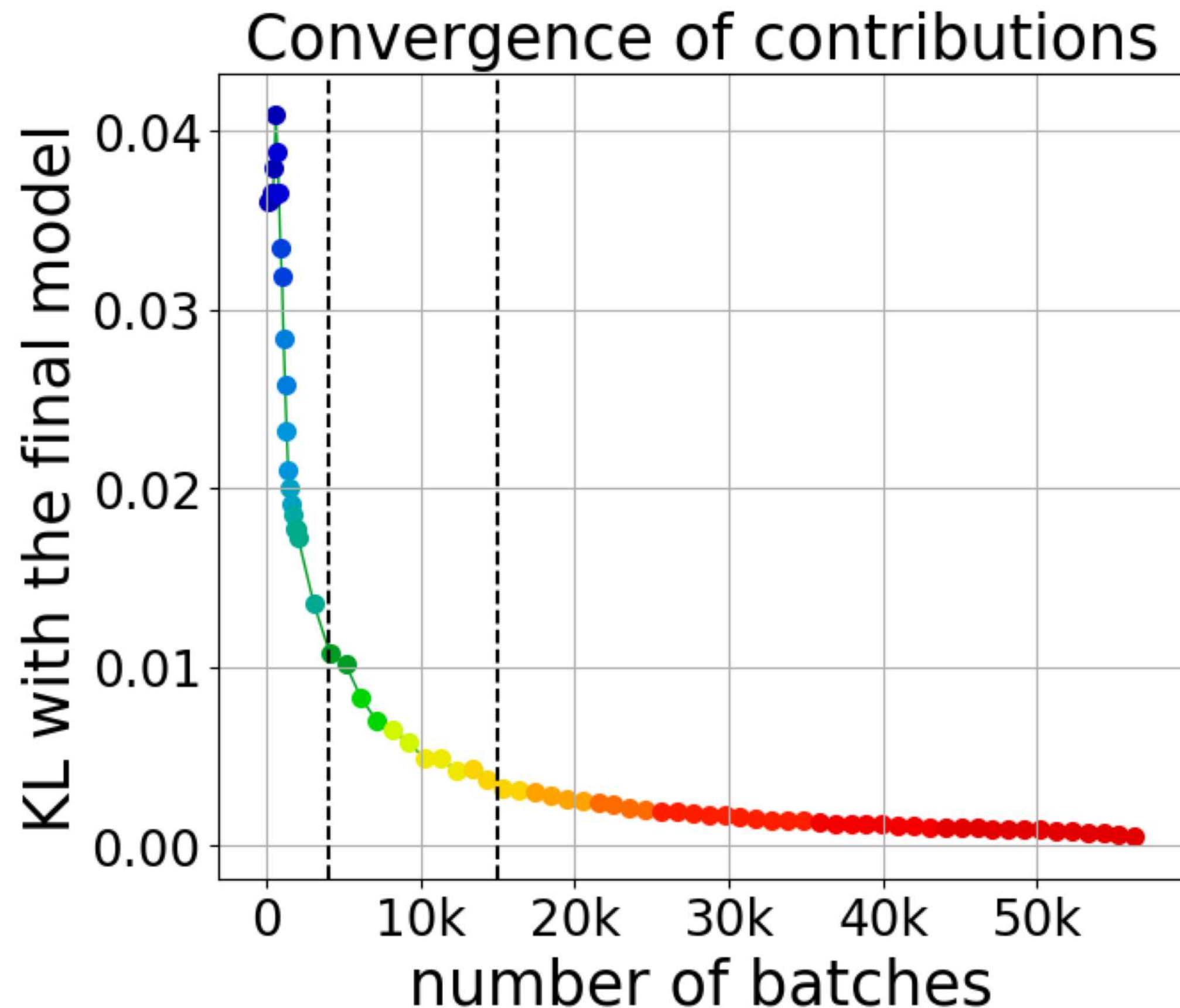
- a lot of change
- more change for early positions
- source influence increases
- entropy of contributions increases

- almost no change
- entropy of contributions slightly decreases





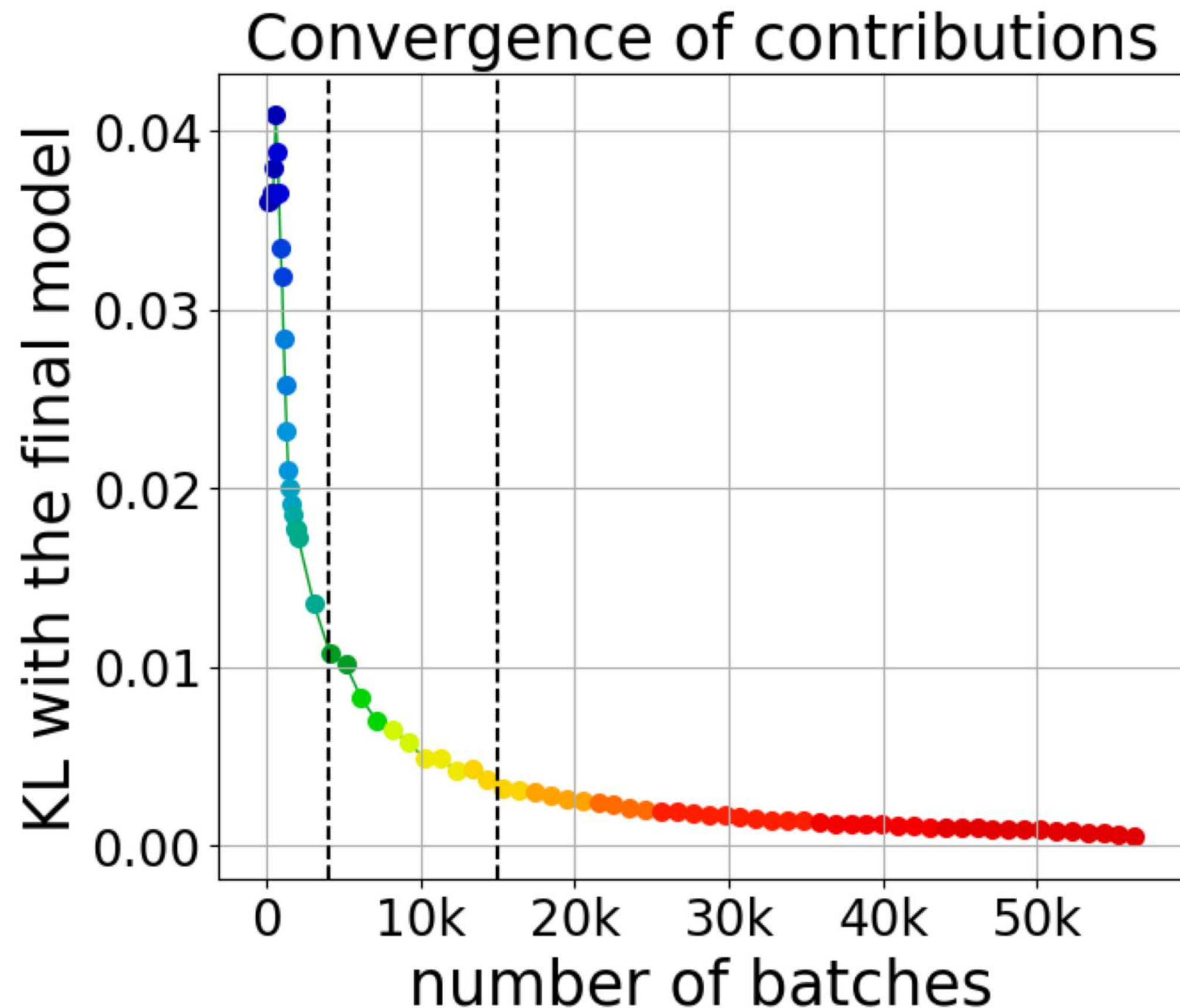
# Changes in Contributions



How:

- evaluate KL divergence in token influence distributions (between final model and in training)

# Changes in Contributions

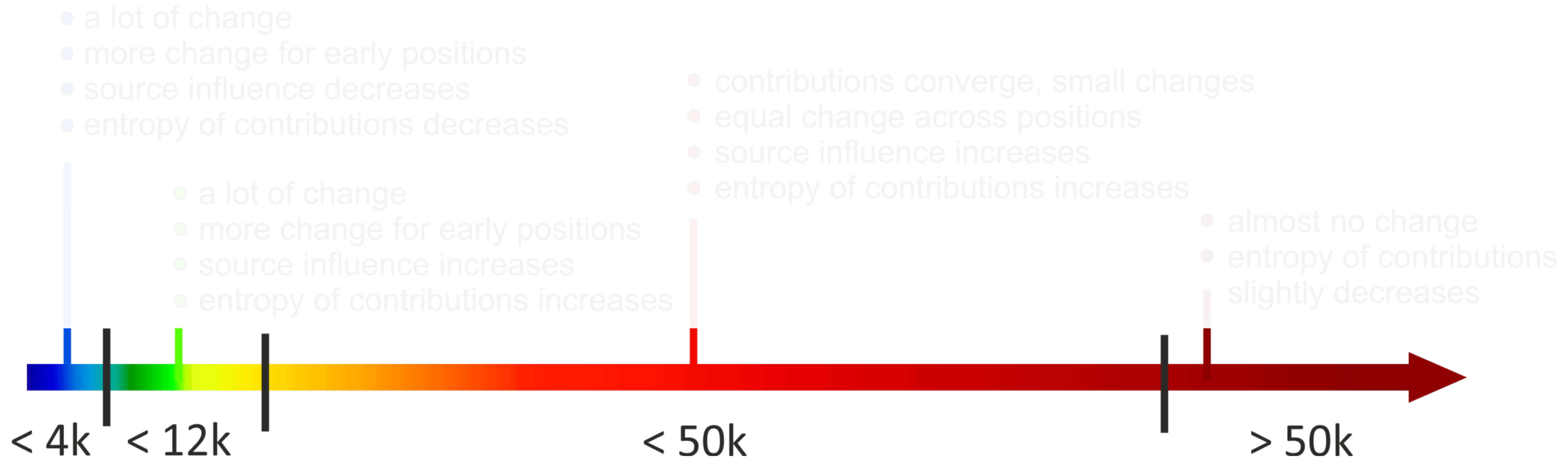


How:

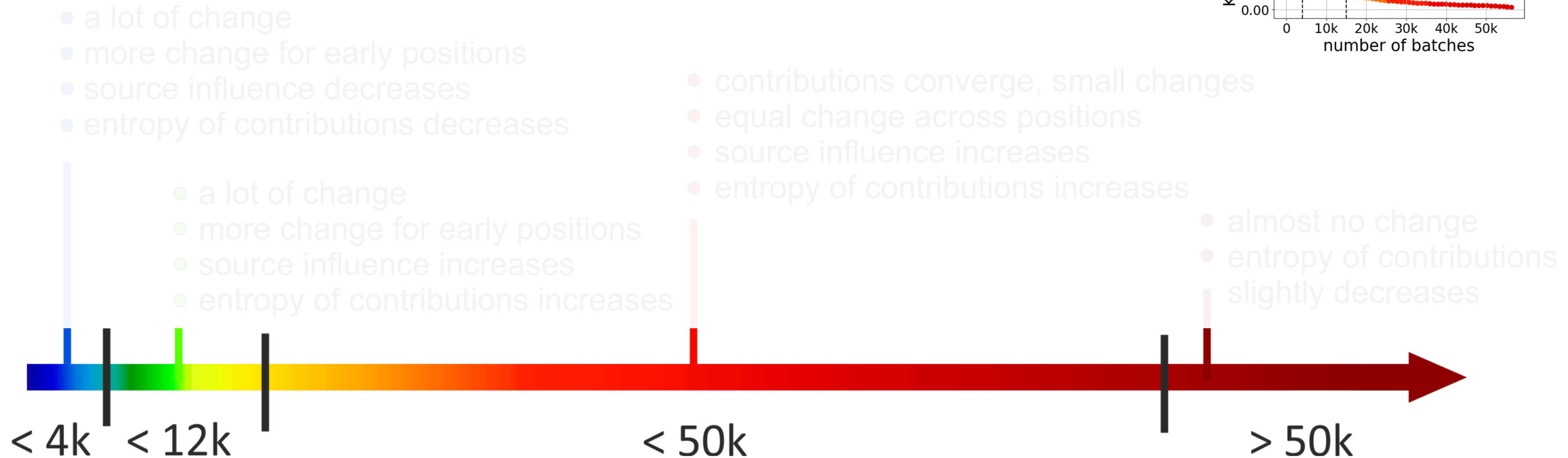
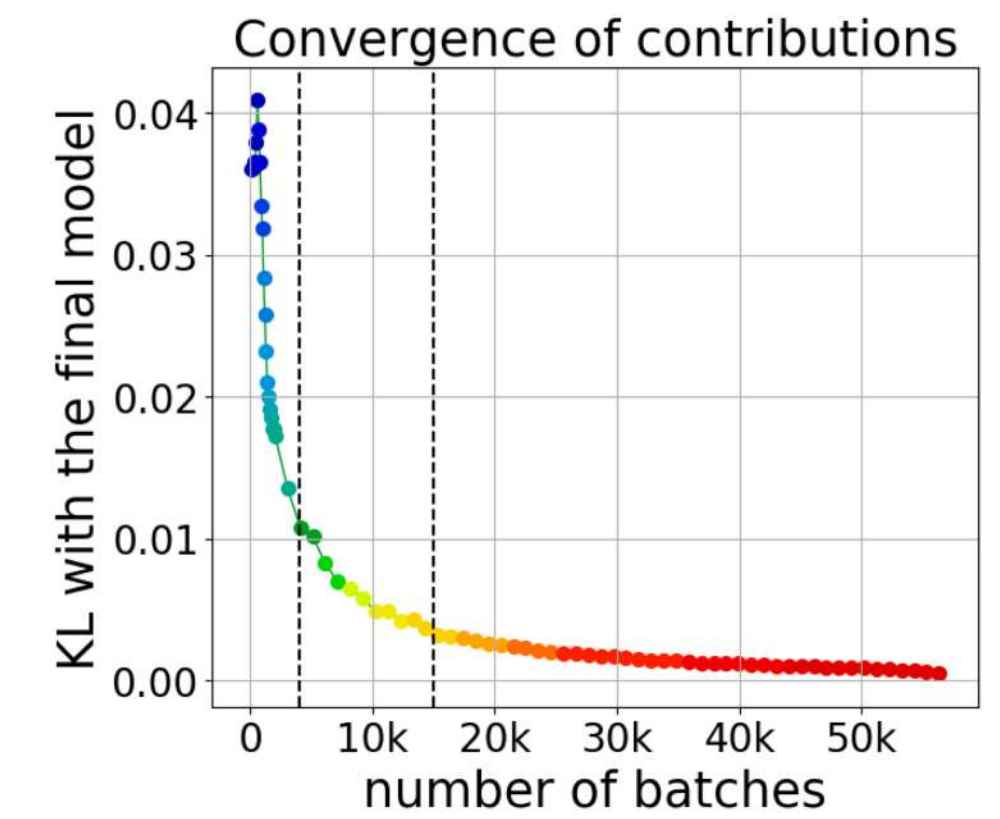
- evaluate KL divergence in token influence distributions (between final model and in training)

Early in training, the model is already close to its final state in the choice of important tokens

# The Training Timeline

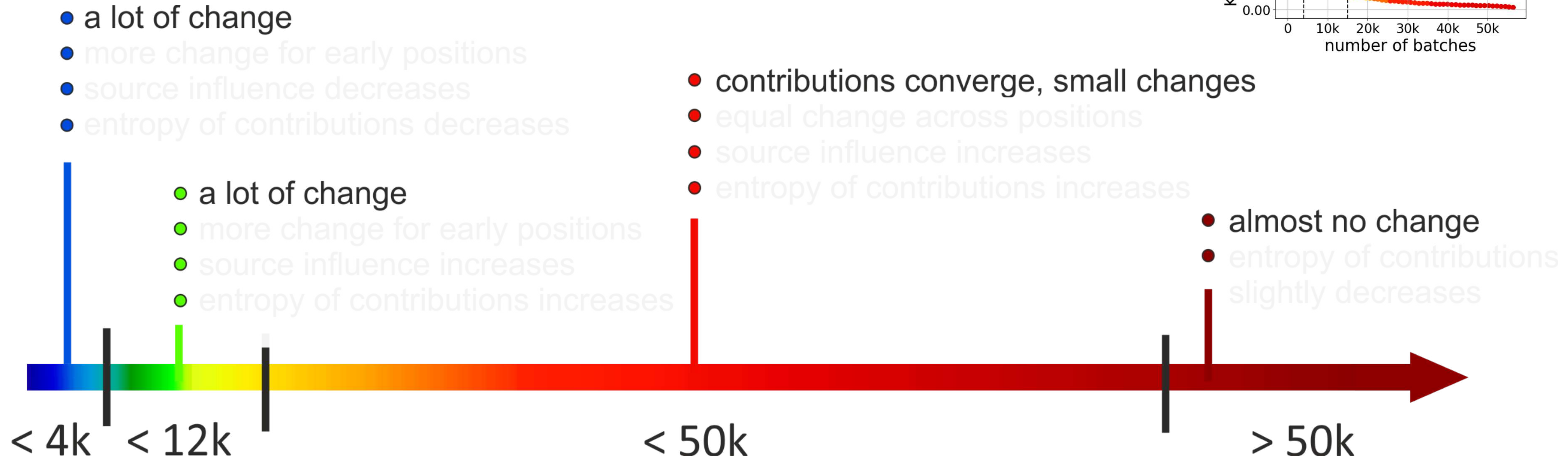
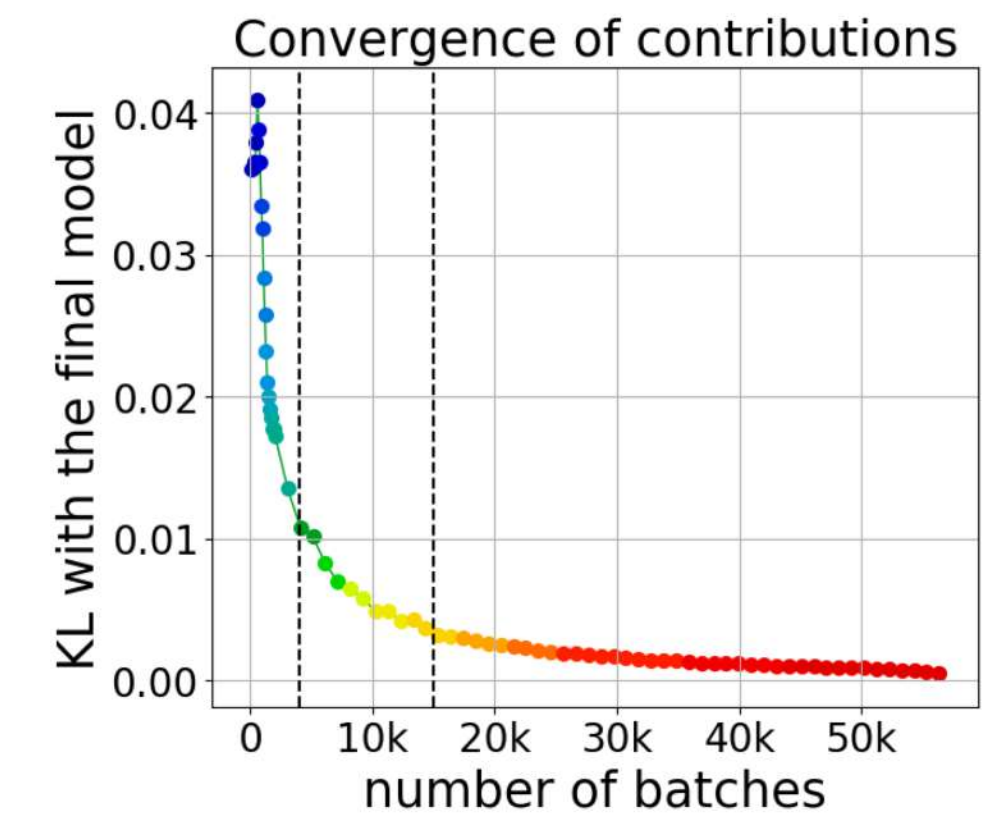


# The Training Timeline



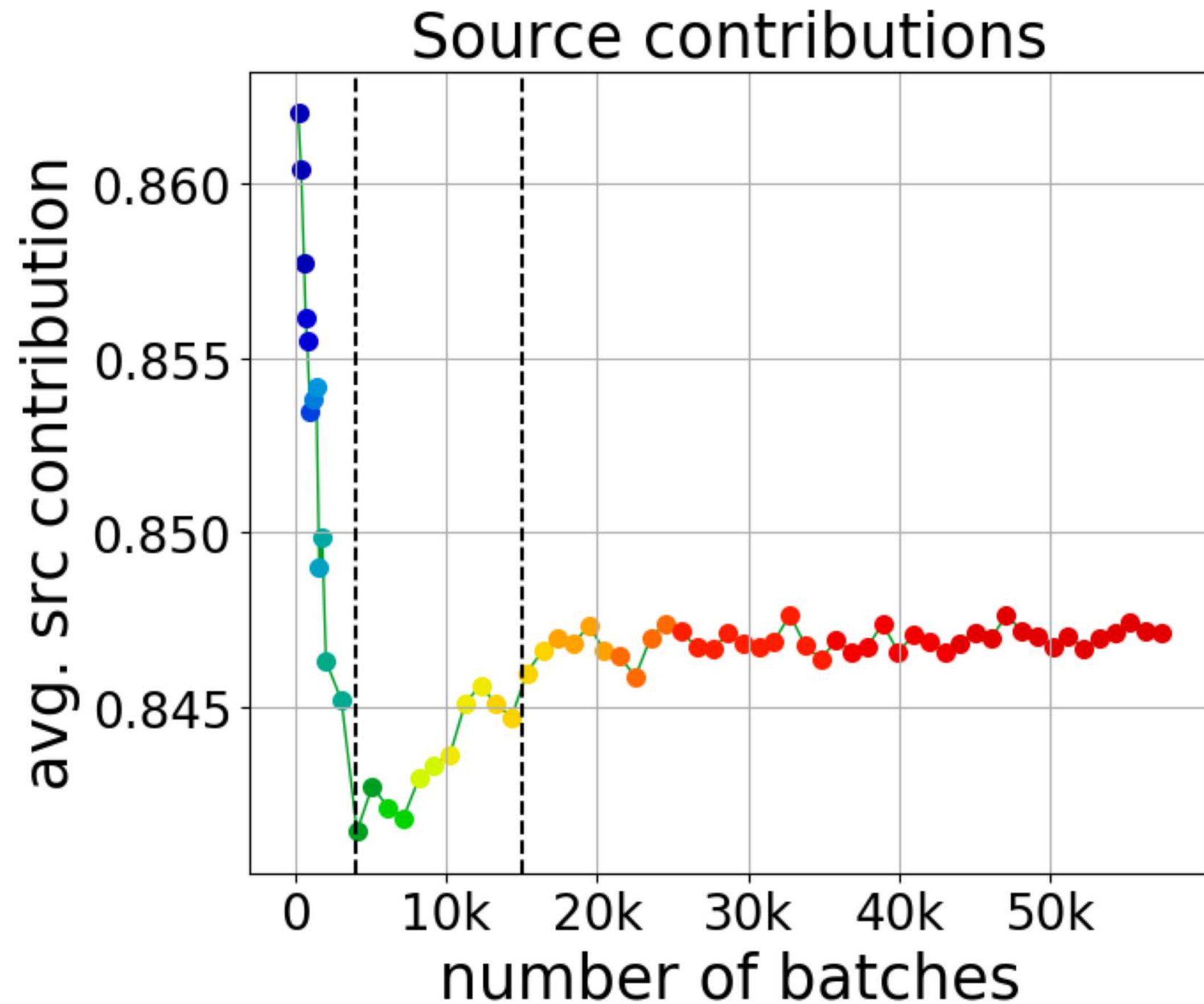


# The Training Timeline





# Source Contributions: Changes are Not Monotonic

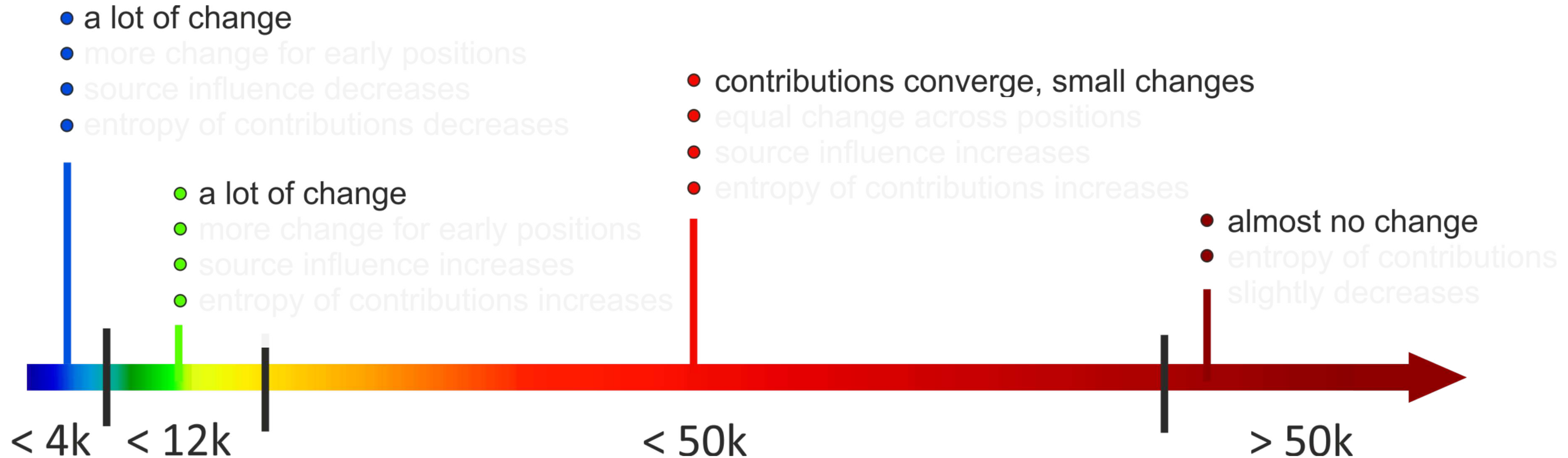


How:

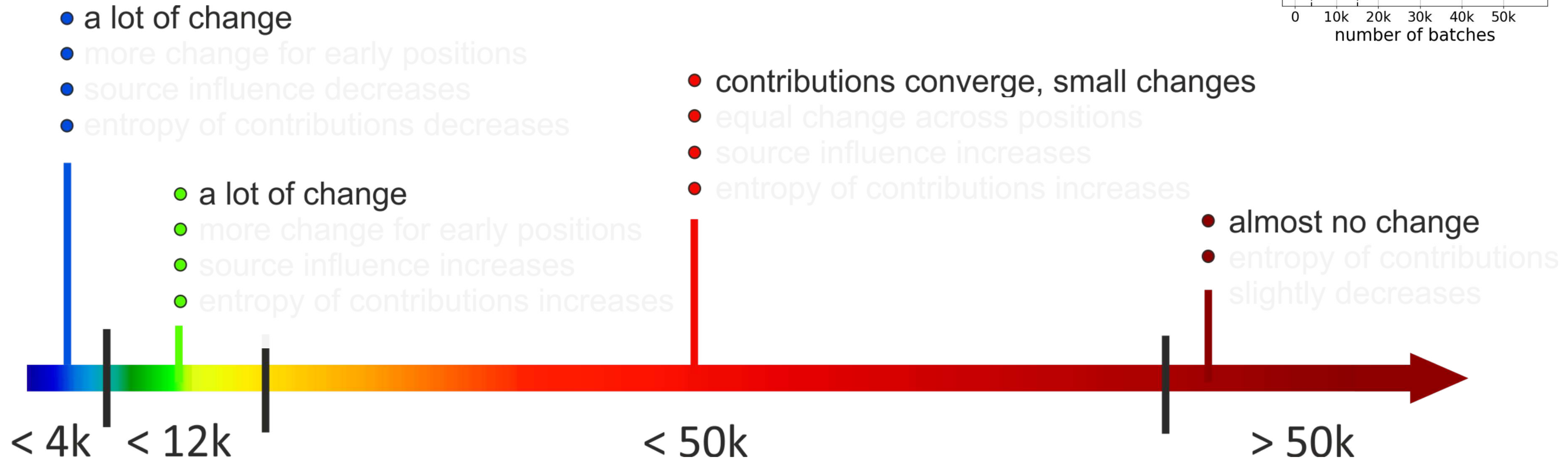
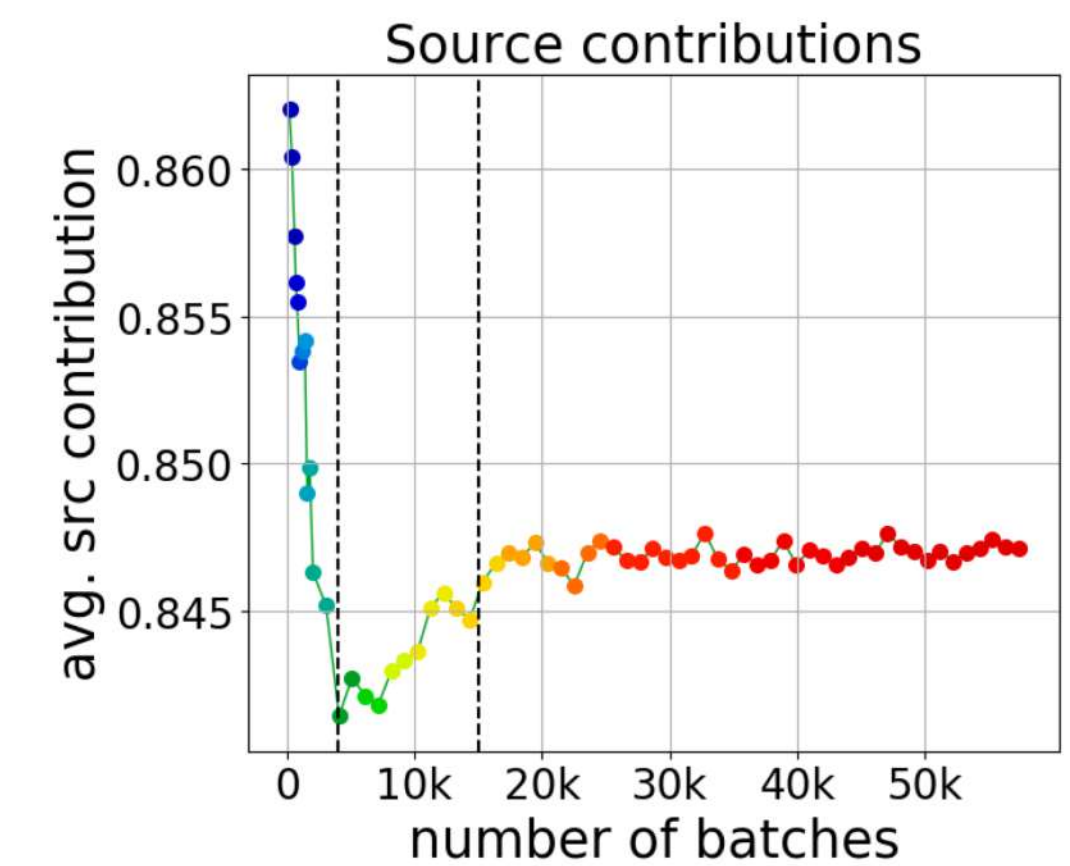
- evaluate average (over target positions and examples) source contribution

Changes are NOT monotonic!

# The Training Timeline

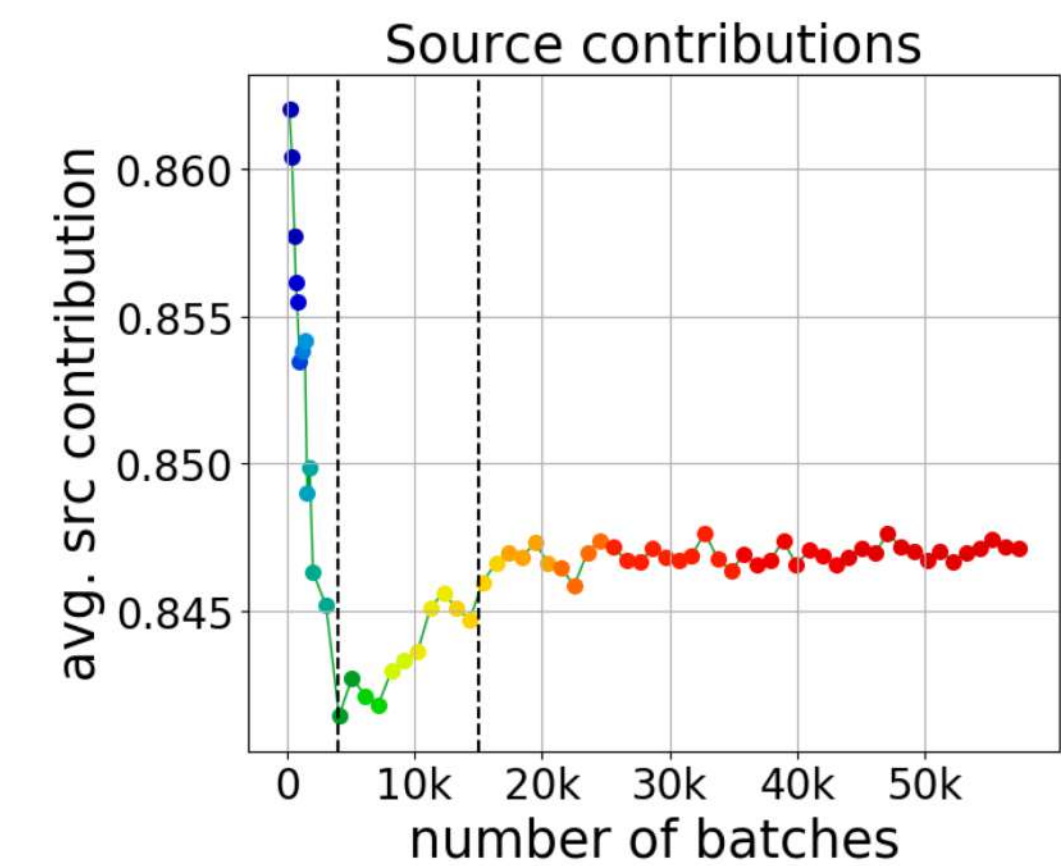


# The Training Timeline





# The Training Timeline

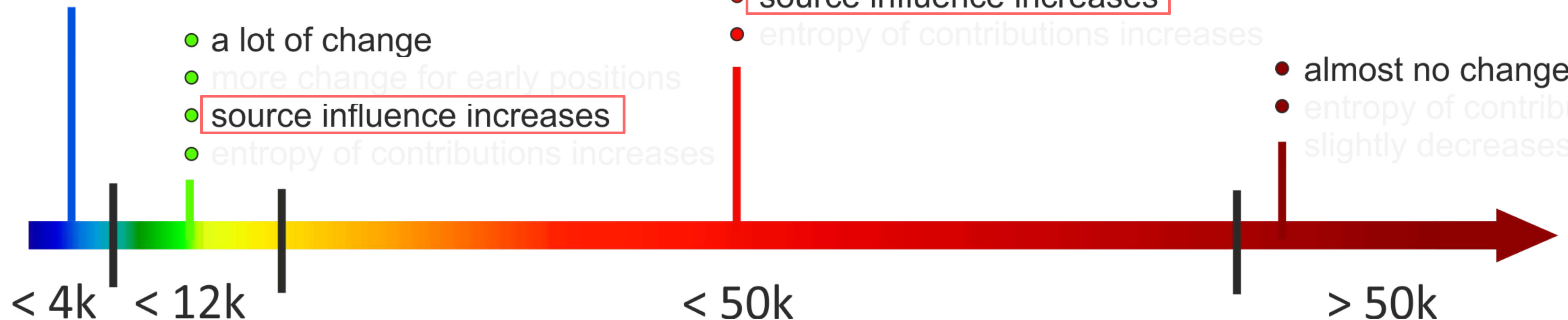


- a lot of change
- more change for early positions
- source influence decreases
- entropy of contributions decreases

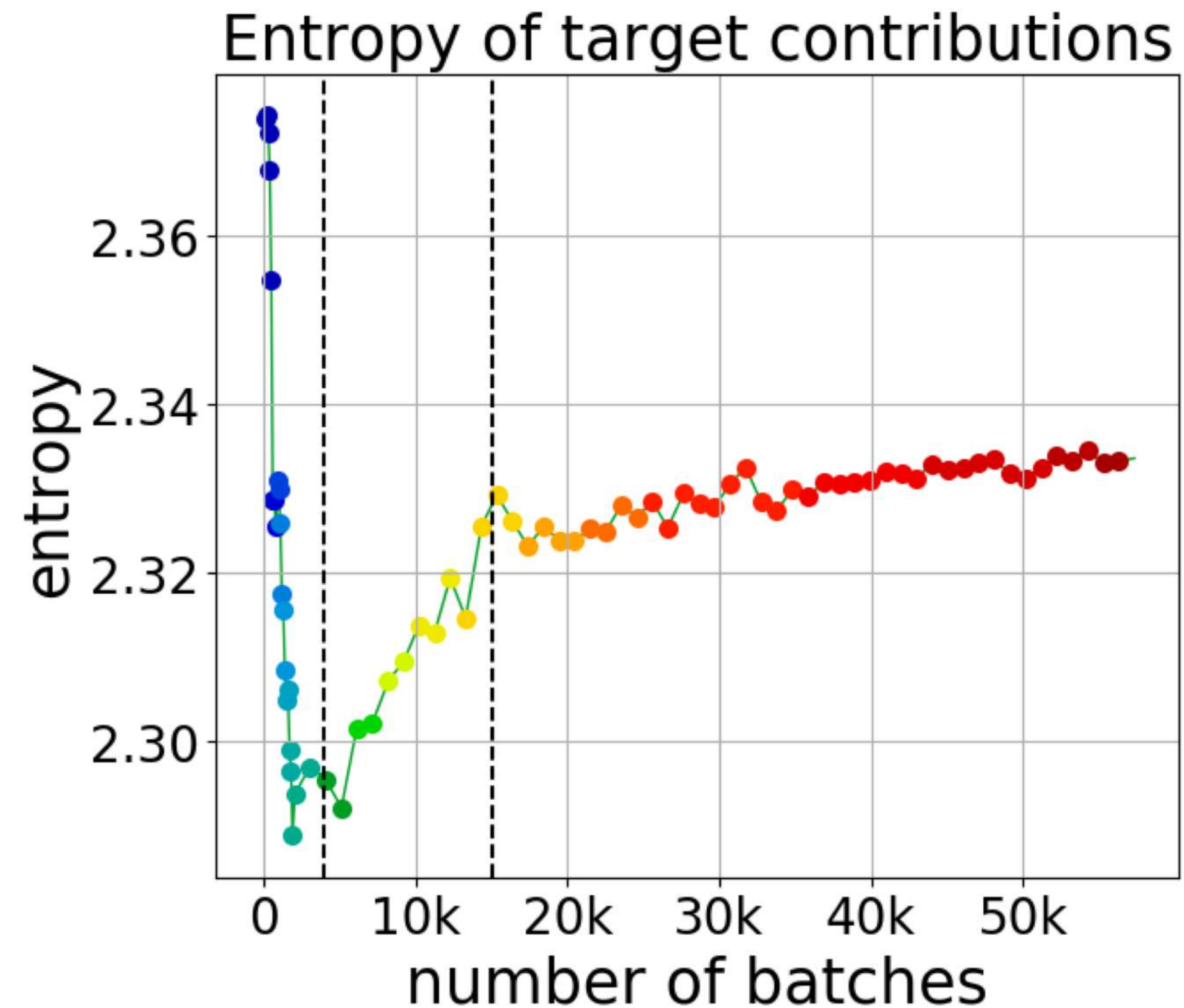
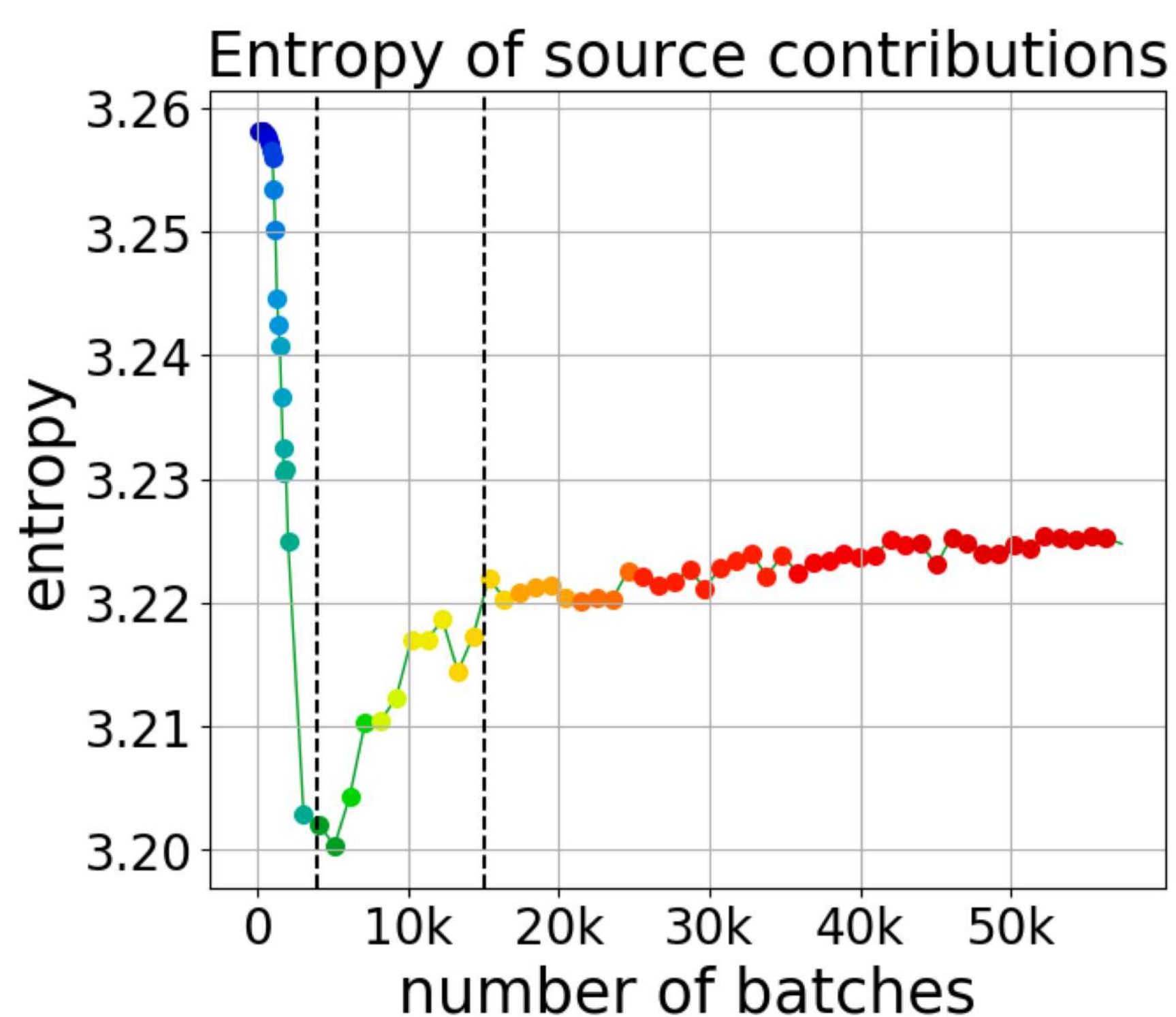
- contributions converge, small changes
- equal change across positions
- source influence increases
- entropy of contributions increases

- a lot of change
- more change for early positions
- source influence increases
- entropy of contributions increases

- almost no change
- entropy of contributions slightly decreases



# Entropy of Contributions: Not Monotonic

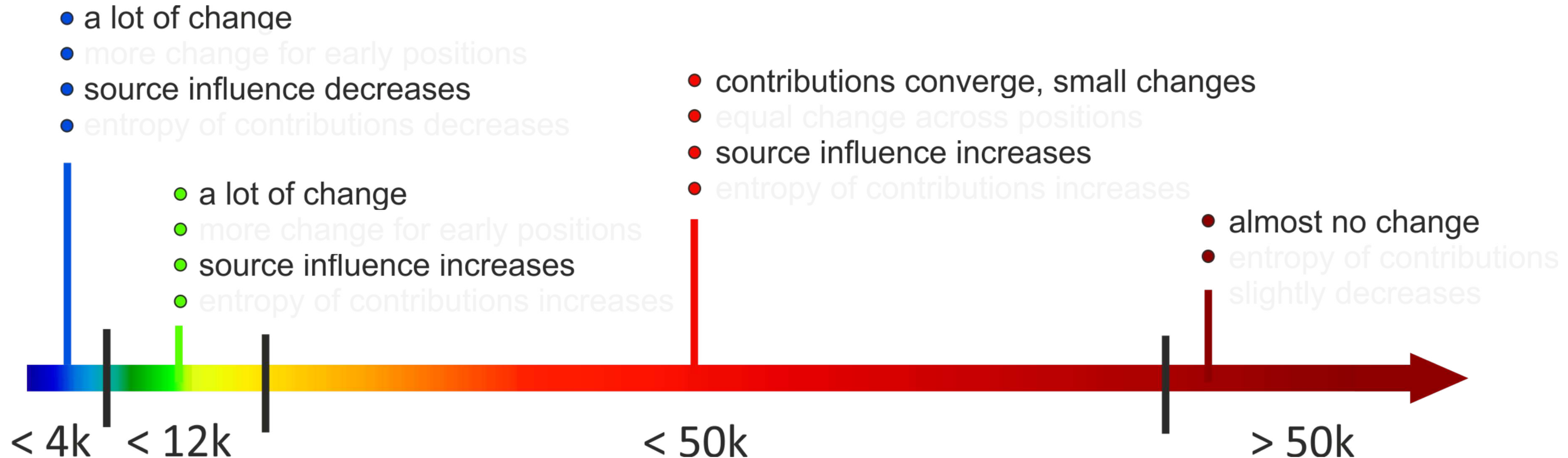


Changes are not monotonic:

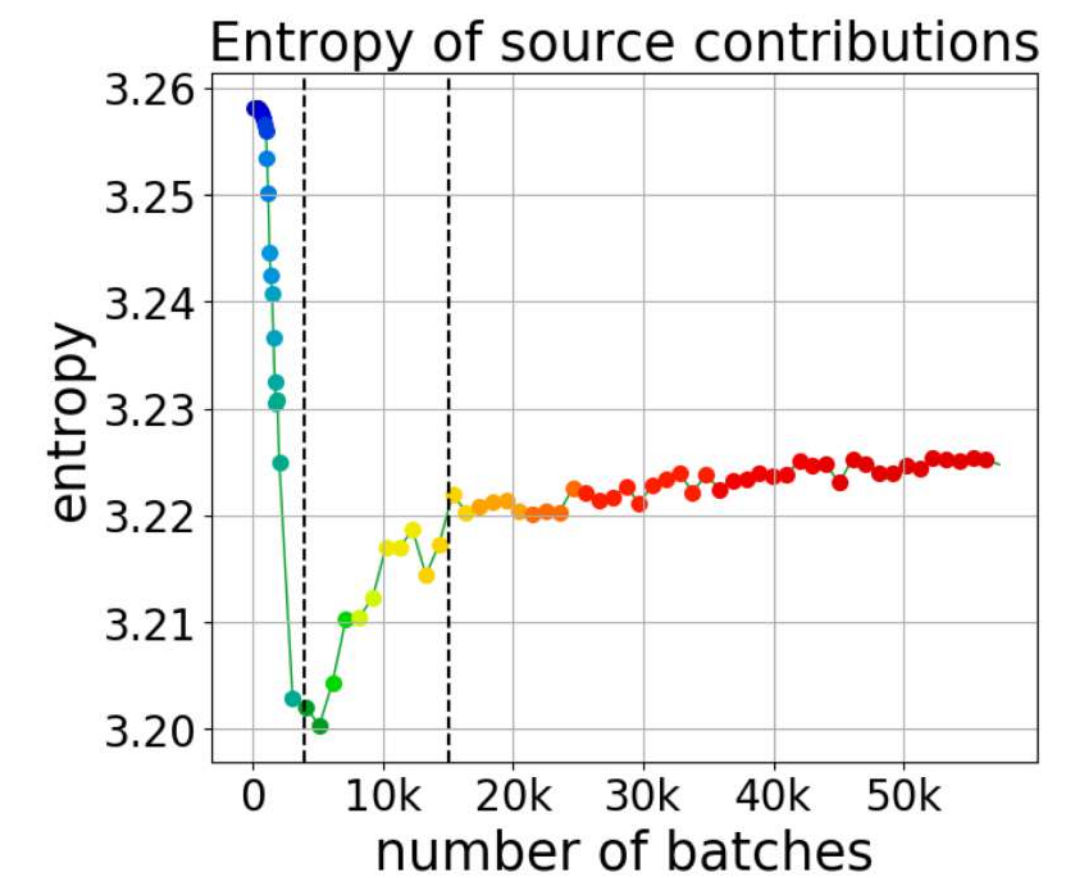
- decrease (maybe learn simple things, e.g., word-by-word translation)
- increase (learn more complex things and rely on broader context)



# The Training Timeline



# The Training Timeline

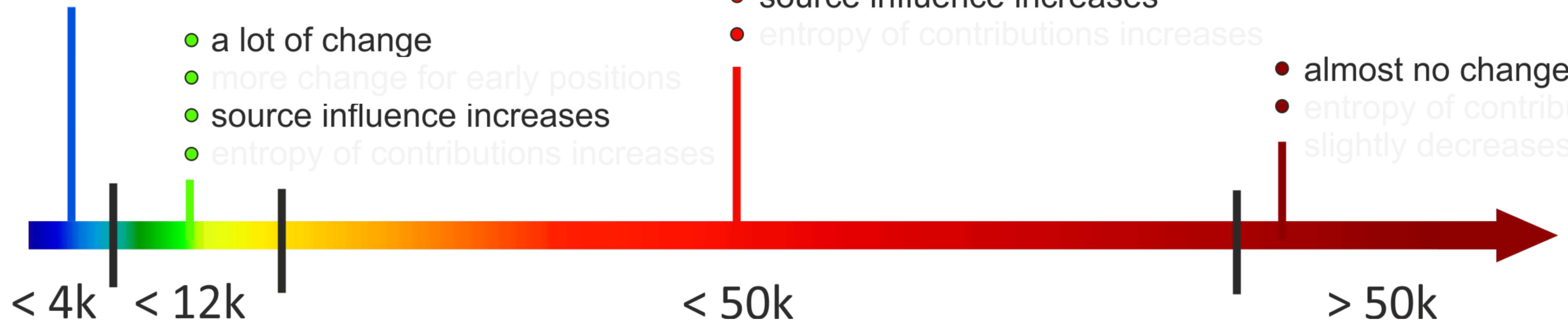


- a lot of change
- more change for early positions
- source influence decreases
- entropy of contributions decreases

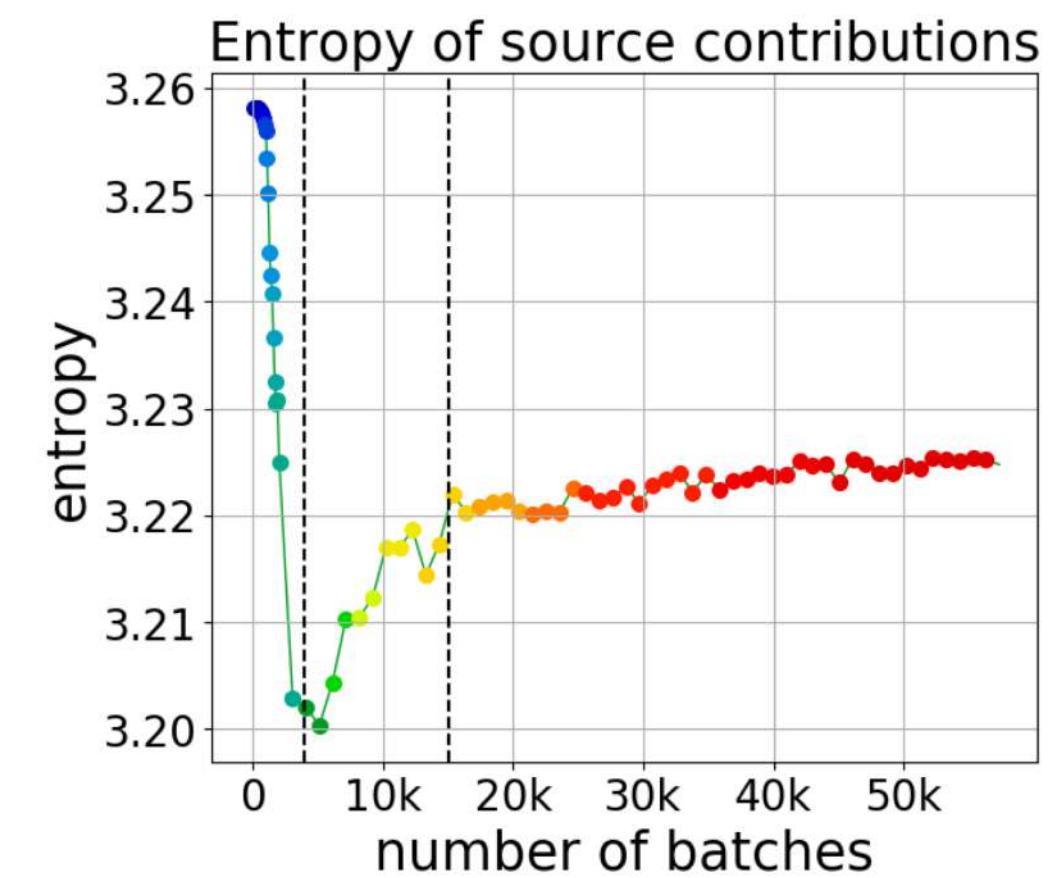
- contributions converge, small changes
- equal change across positions
- source influence increases
- entropy of contributions increases

- a lot of change
- more change for early positions
- source influence increases
- entropy of contributions increases

- almost no change
- entropy of contributions slightly decreases



# The Training Timeline

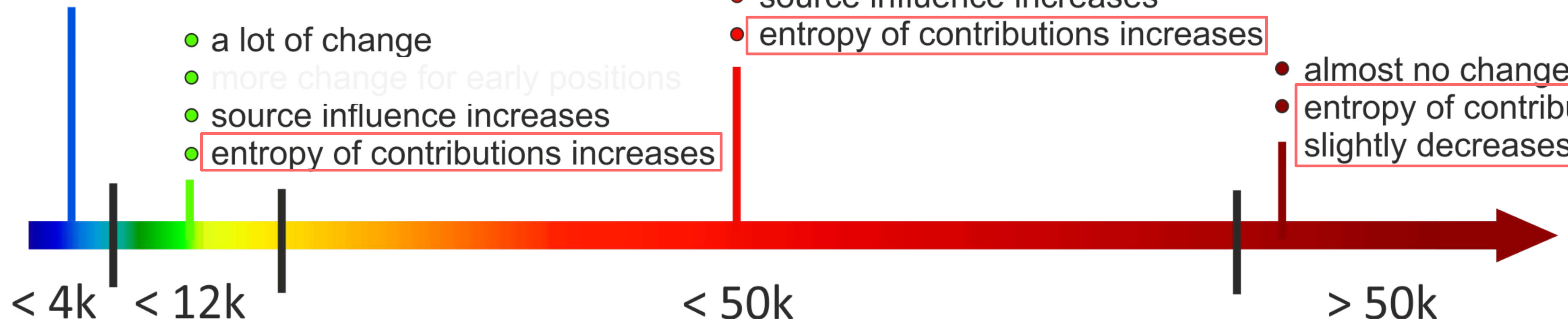


- a lot of change
- more change for early positions
- source influence decreases
- entropy of contributions decreases

- a lot of change
- more change for early positions
- source influence increases
- entropy of contributions increases

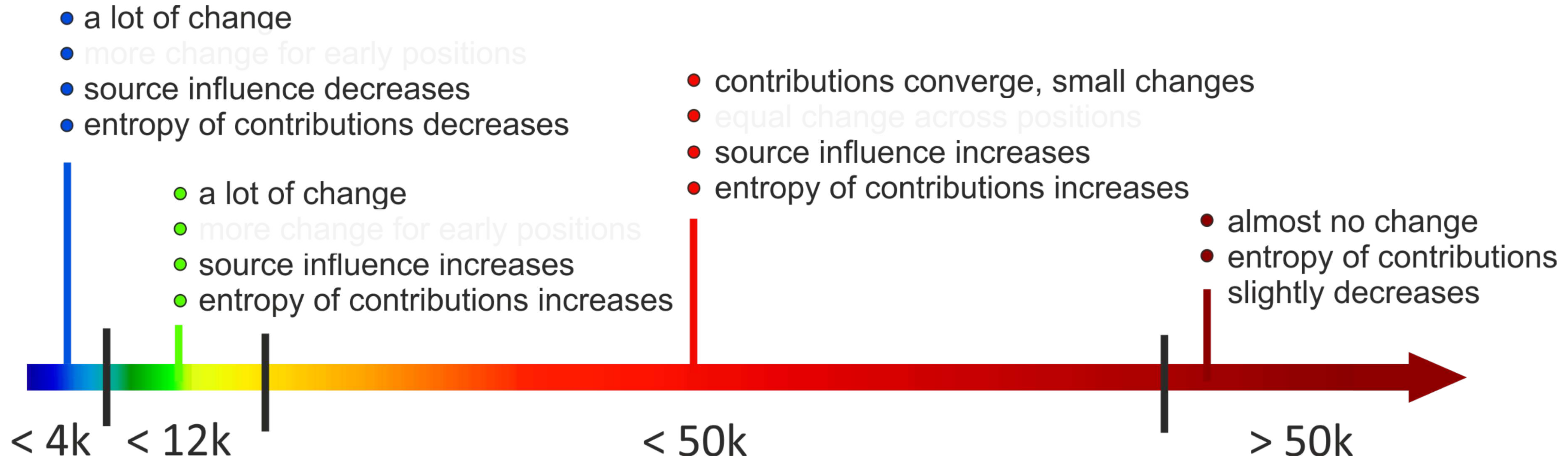
- contributions converge, small changes
- equal change across positions
- source influence increases
- entropy of contributions increases

- almost no change
- entropy of contributions slightly decreases

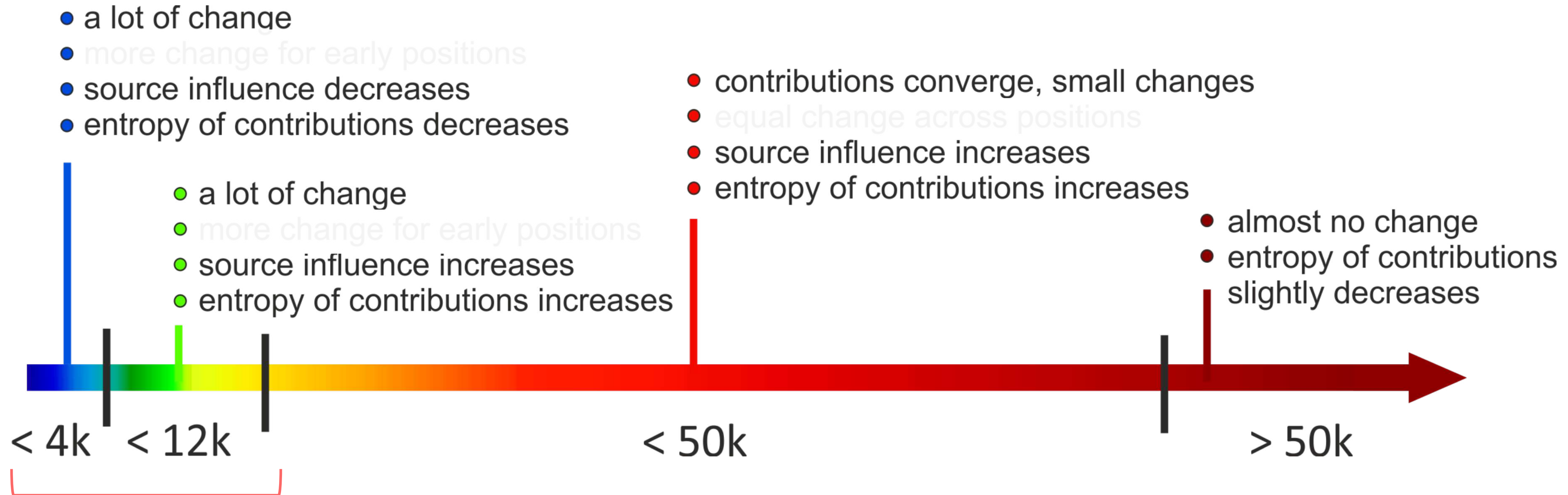




# The Training Timeline



# The Training Timeline



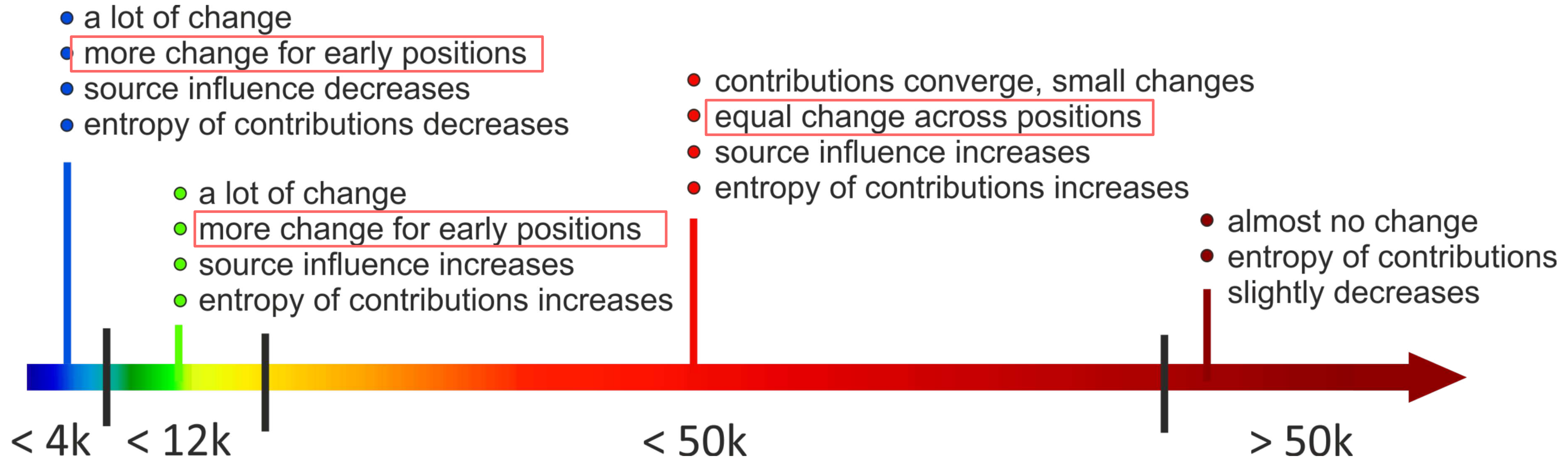
Early positions change more and are learned first



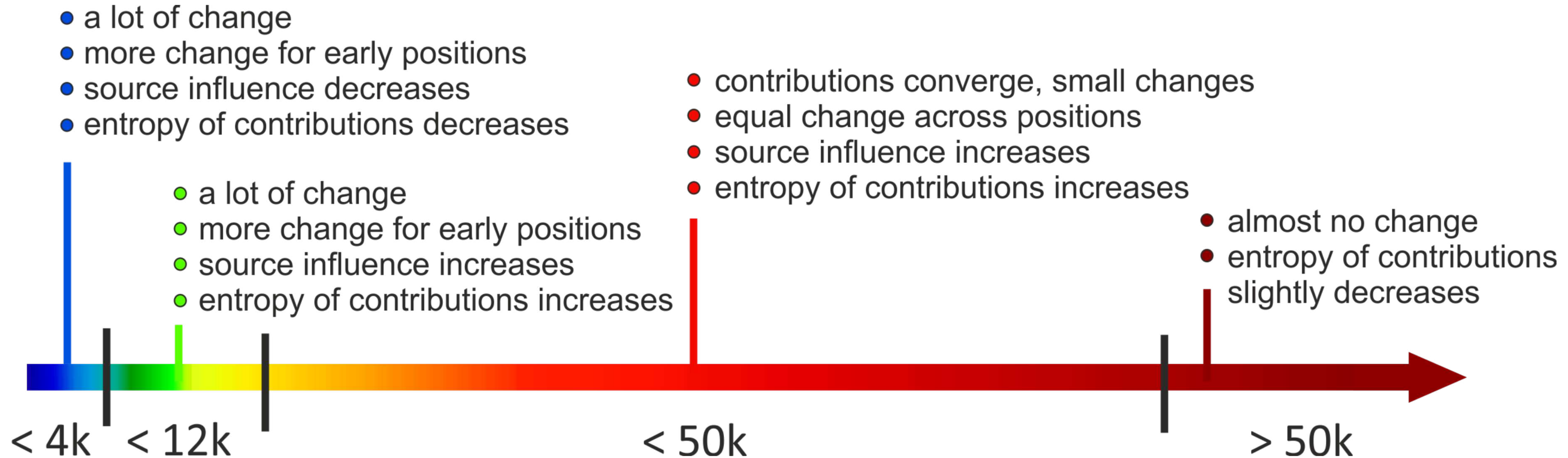
More details in the paper!



# The Training Timeline

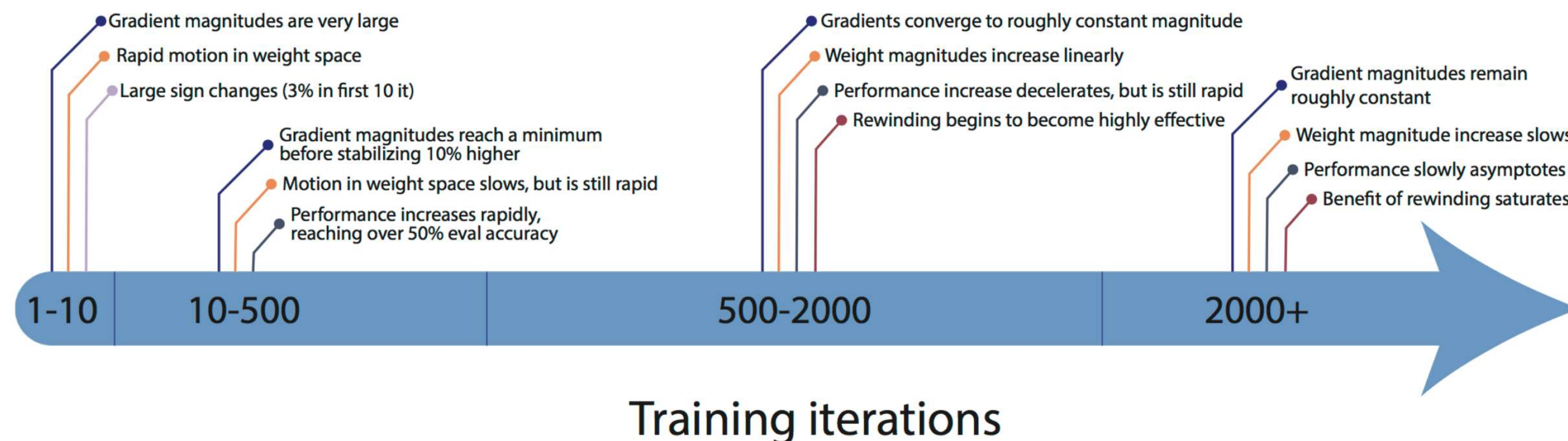


# The Training Timeline

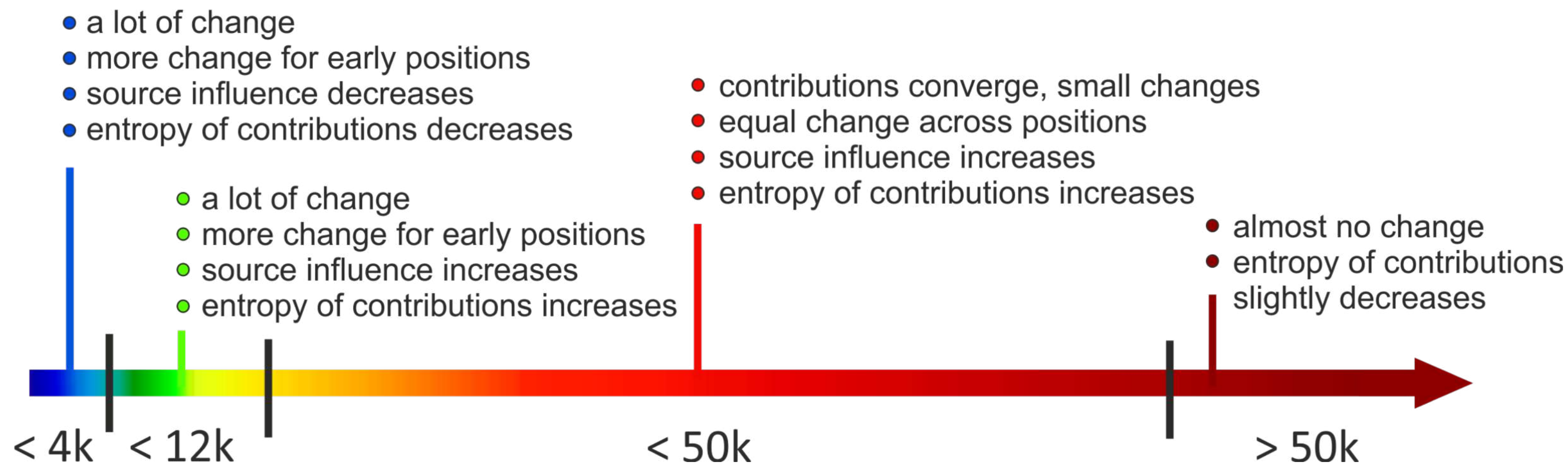




# Relation to Previous Work

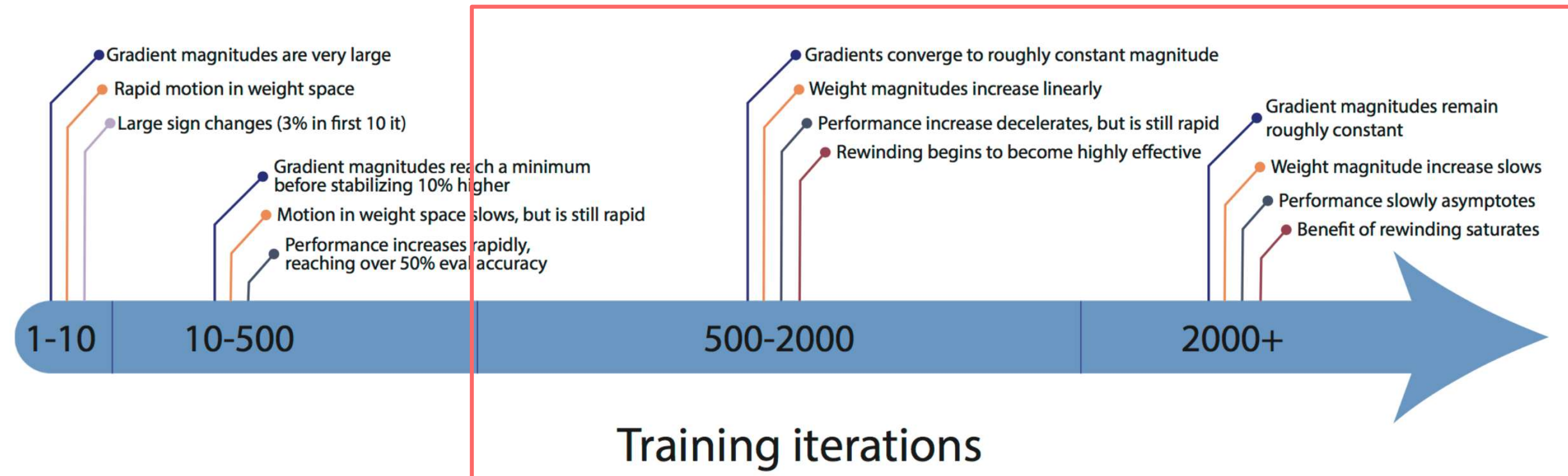


ResNet-20 on CIFAR-10, Frankle et al, ICLR 2020

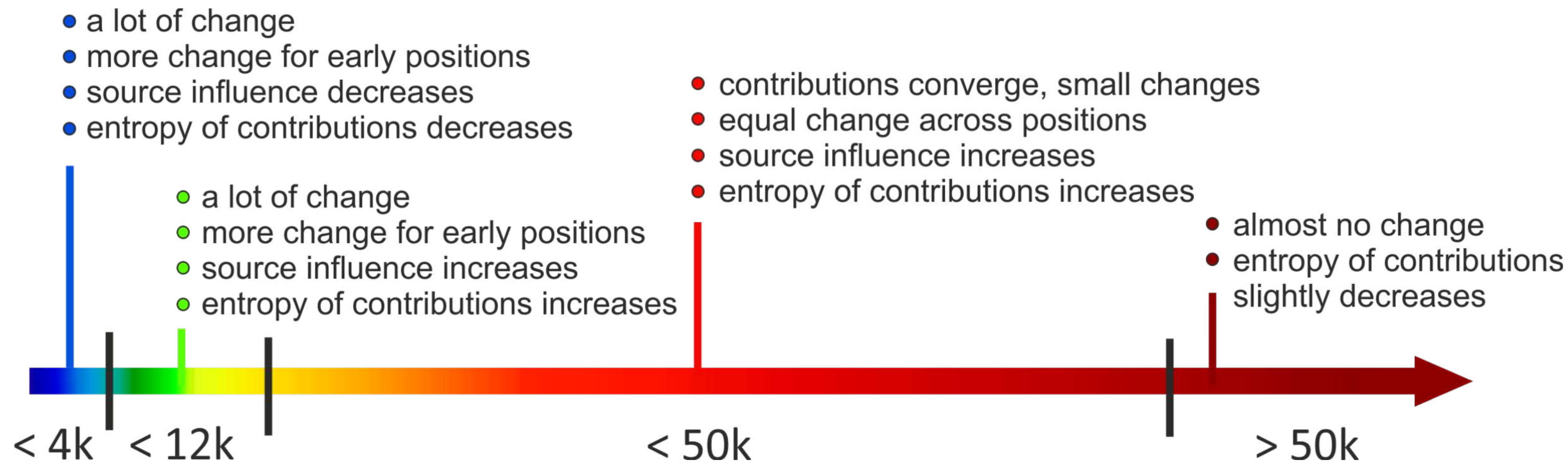




# Relation to Previous Work



ResNet-20 on CIFAR-10, Frankle et al, ICLR 2020



Stages by Frankle et al, 2020:

- found when validating the lottery ticket (LT) hypothesis
- match well with ours
- rewinding (for LT) starts to work at stage 3 - when the contributions already converged

# What is going to happen:

## The Trade-Off Between Source and Target

- Our Approach: (a version of) LRP
- Experiments
  - Getting Acquainted
  - Reference, Model and Random Prefixes
  - Exposure Bias and Source Contribution
  - Varying the Amount of Data
  - Training Stages

(A Bit of) the Training Process (work in progress)



# What is going to happen:

## The Trade-Off Between Source and Target

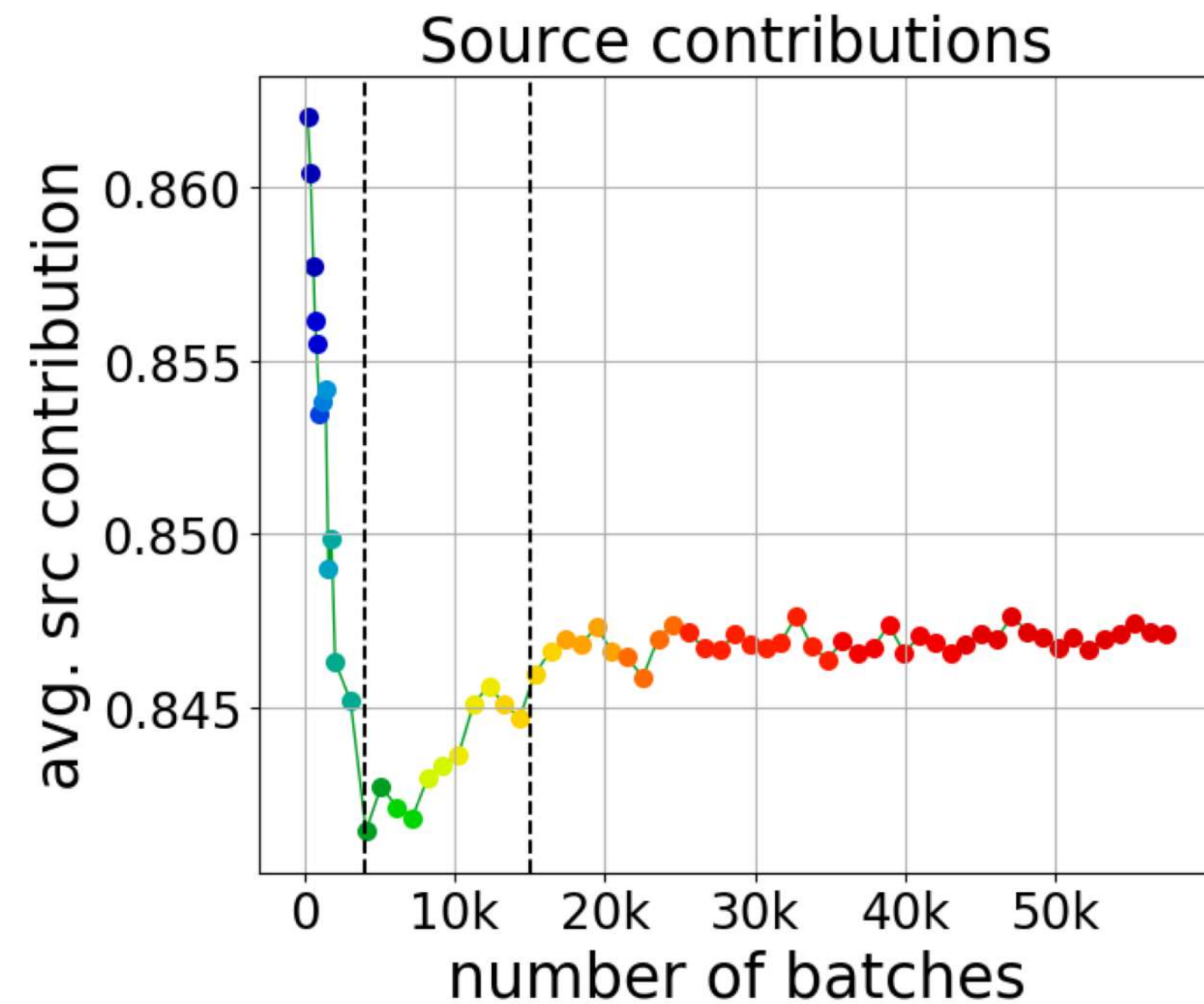
- Our Approach: (a version of) LRP
- Experiments
  - Getting Acquainted
  - Reference, Model and Random Prefixes
  - Exposure Bias and Source Contribution
  - Varying the Amount of Data
  - Training Stages

(A Bit of) the Training Process (work in progress)

# (A Bit More of) The Training Process



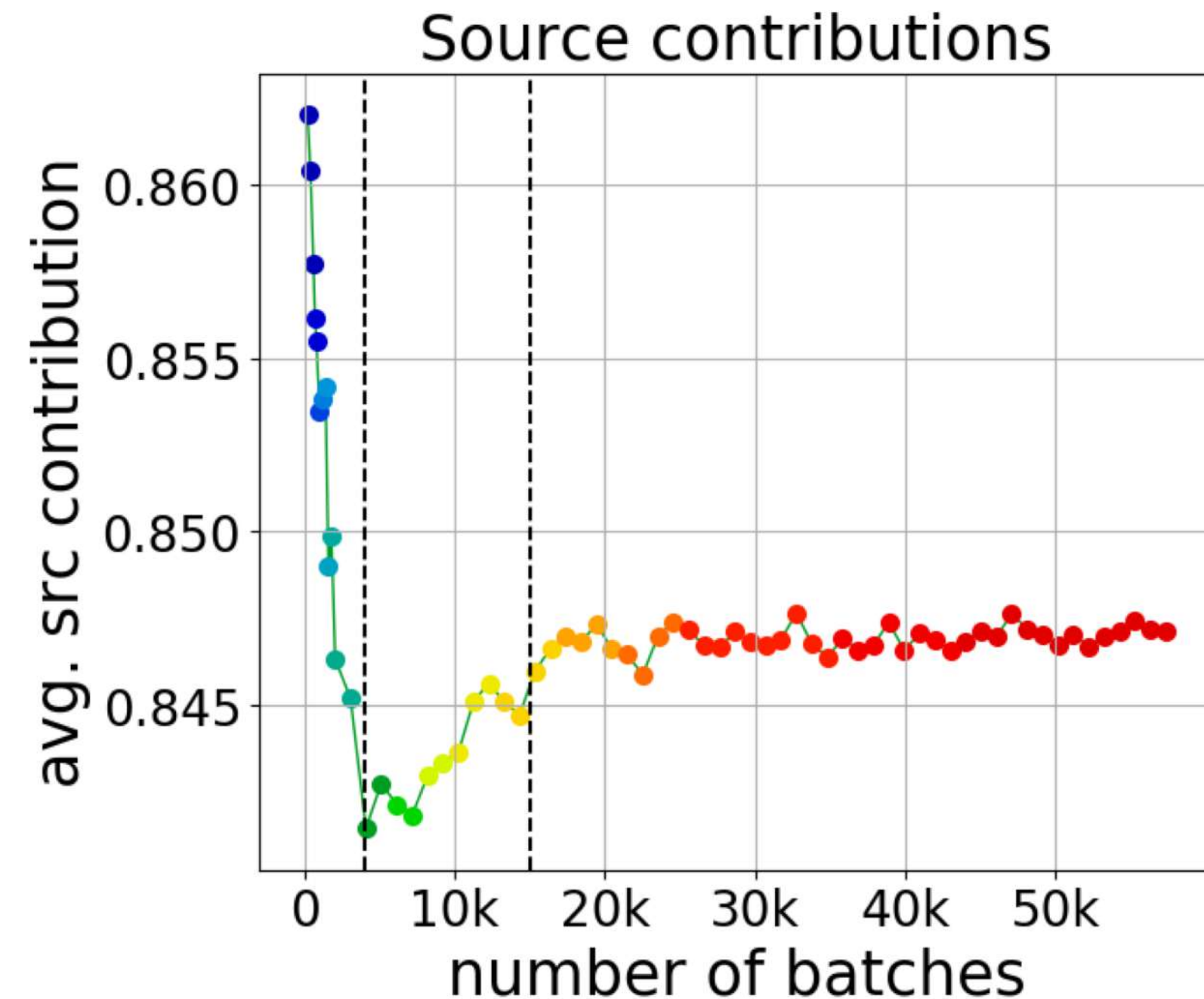
# Contributions and their Entropy: Not Monotonic



(entropy behaves similarly)

# Contributions and their Entropy: Not Monotonic

Changes are not monotonic:

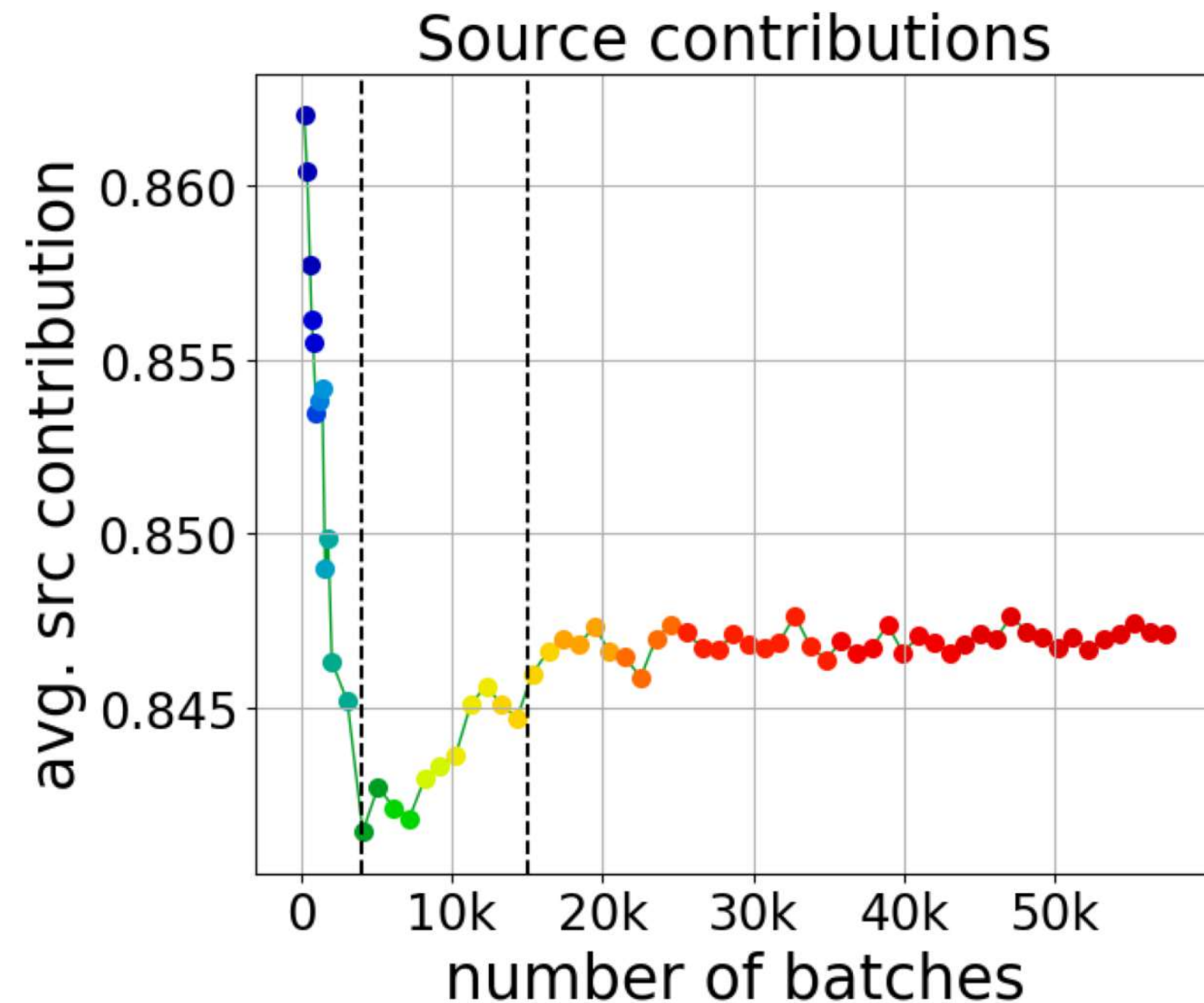


- decrease (maybe learn simple things, e.g., word-by-word translation)

(entropy behaves similarly)

# Contributions and their Entropy: Not Monotonic

Changes are not monotonic:



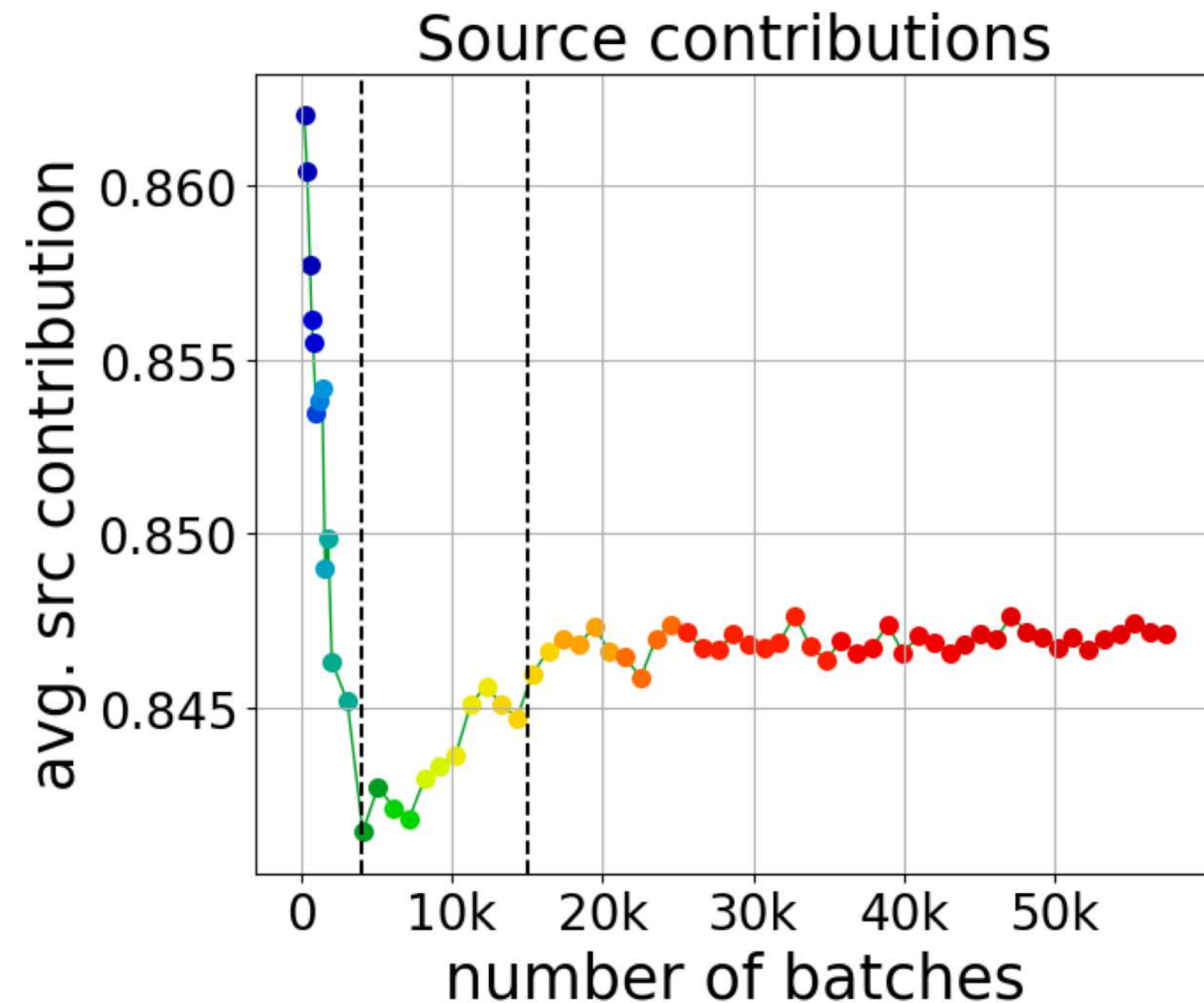
(entropy behaves similarly)

- decrease (maybe learn simple things, e.g., word-by-word translation)
- increase (learn more complex things and rely on broader context)



# Contributions and their Entropy: Not Monotonic

Changes are not monotonic:

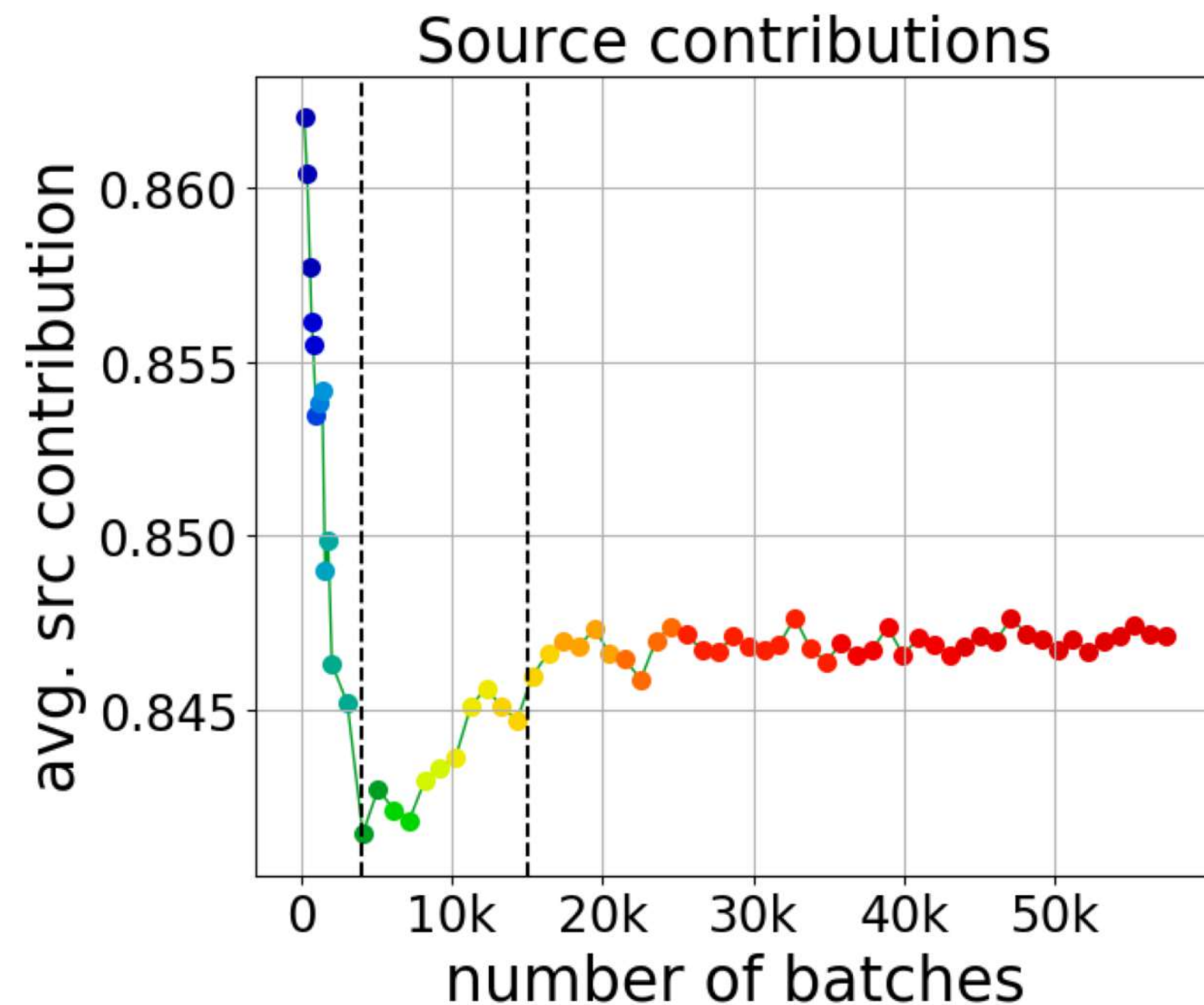


(entropy behaves similarly)

- decrease (maybe learn simple things, e.g., word-by-word translation)
- increase (learn more complex things and rely on broader context)

So far, we only hypothesized what's going on with the model

# Contributions and their Entropy: Not Monotonic

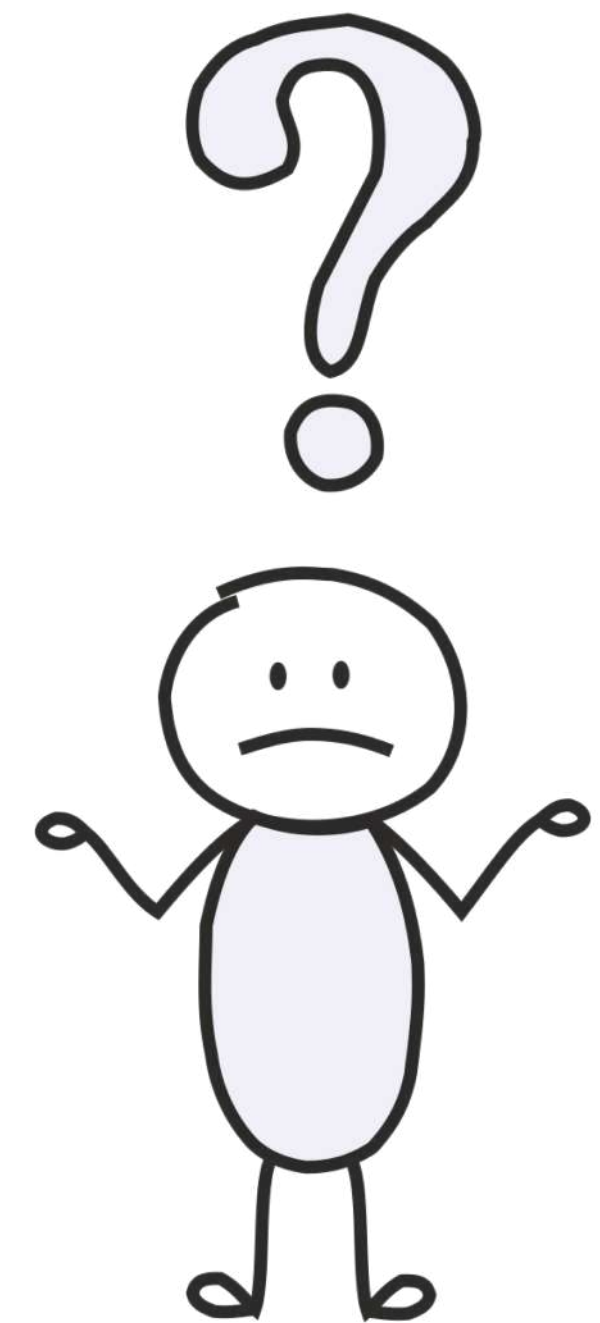


(entropy behaves similarly)

Changes are not monotonic:

- decrease (maybe learn simple things, e.g., word-by-word translation)
- increase (learn more complex things and rely on broader context)

So far, we only hypothesized what's going on with the model



But what is **really** going on?

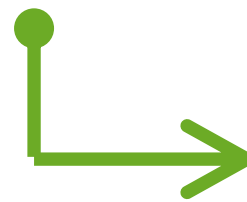
# Take a Step Back: SMT vs NMT

Statistical MT

Neural MT

# Take a Step Back: SMT vs NMT

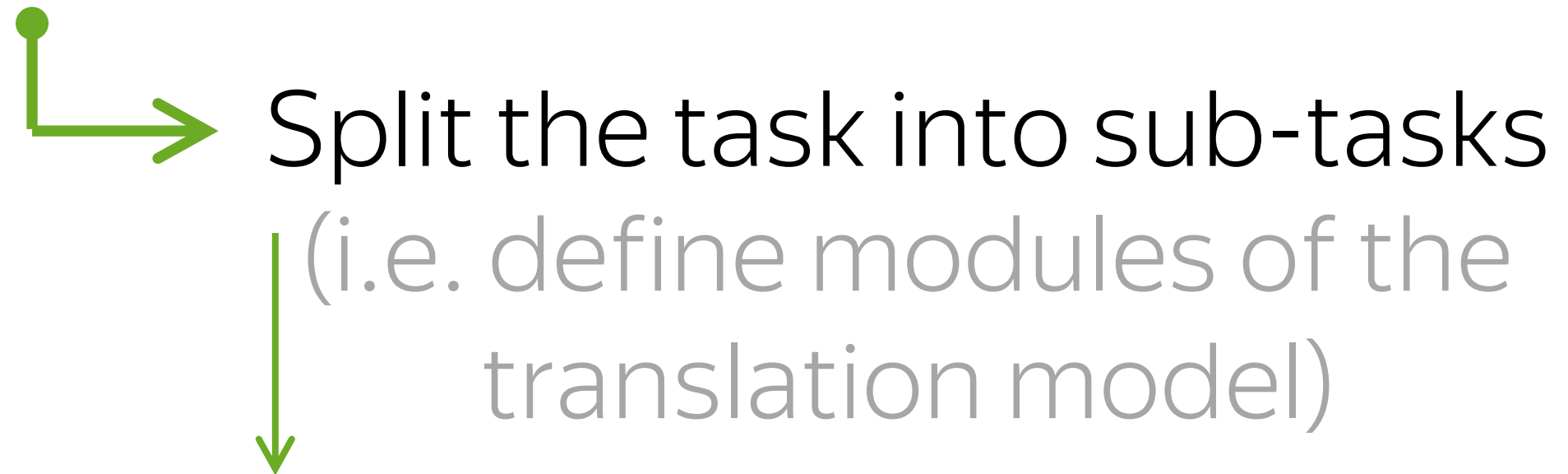
Statistical MT

 Split the task into sub-tasks  
(i.e. define modules of the  
translation model)

Neural MT

# Take a Step Back: SMT vs NMT

## Statistical MT



Train these modules

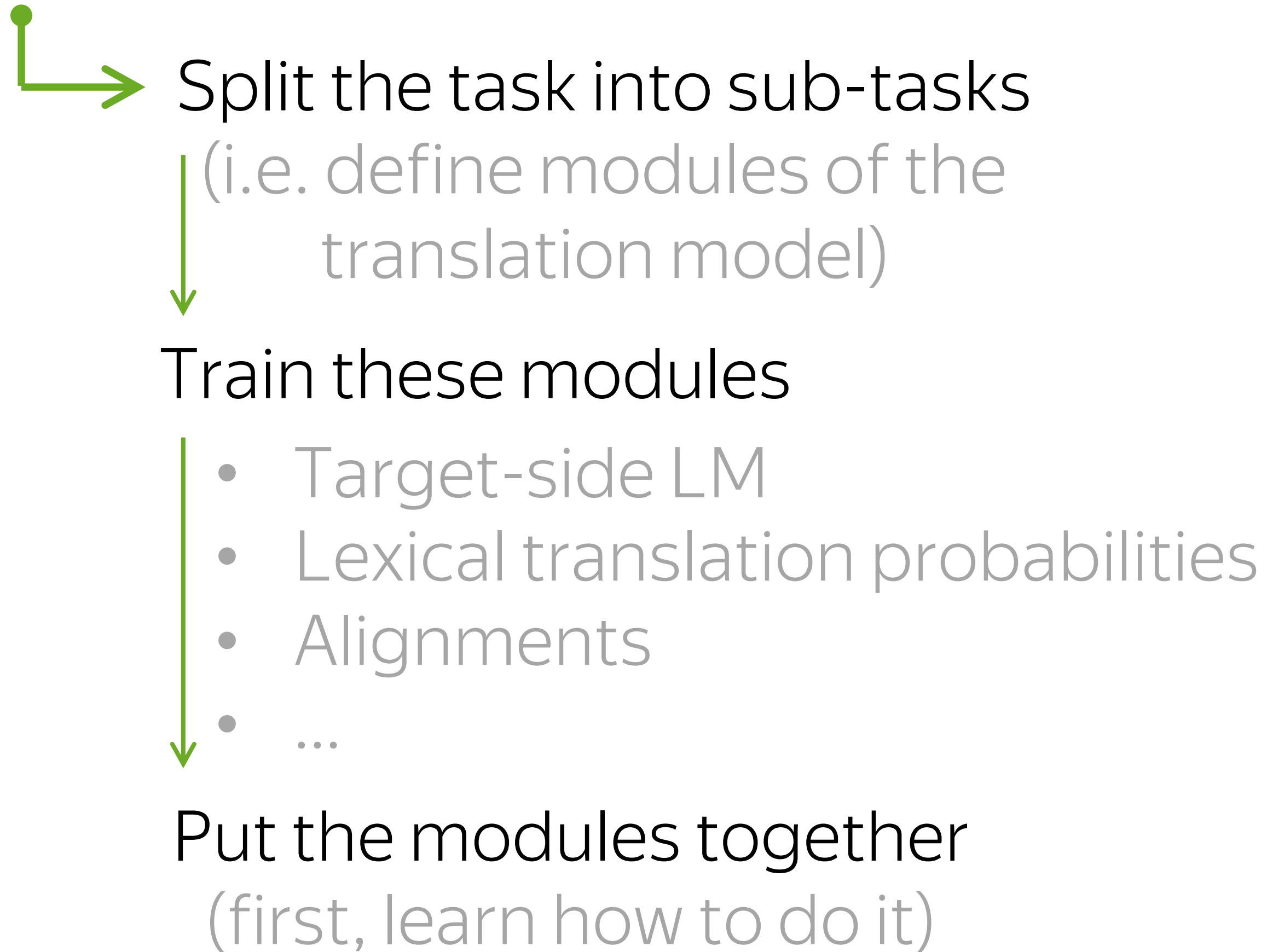
- Target-side LM
- Lexical translation probabilities
- Alignments
- ...

## Neural MT



# Take a Step Back: SMT vs NMT

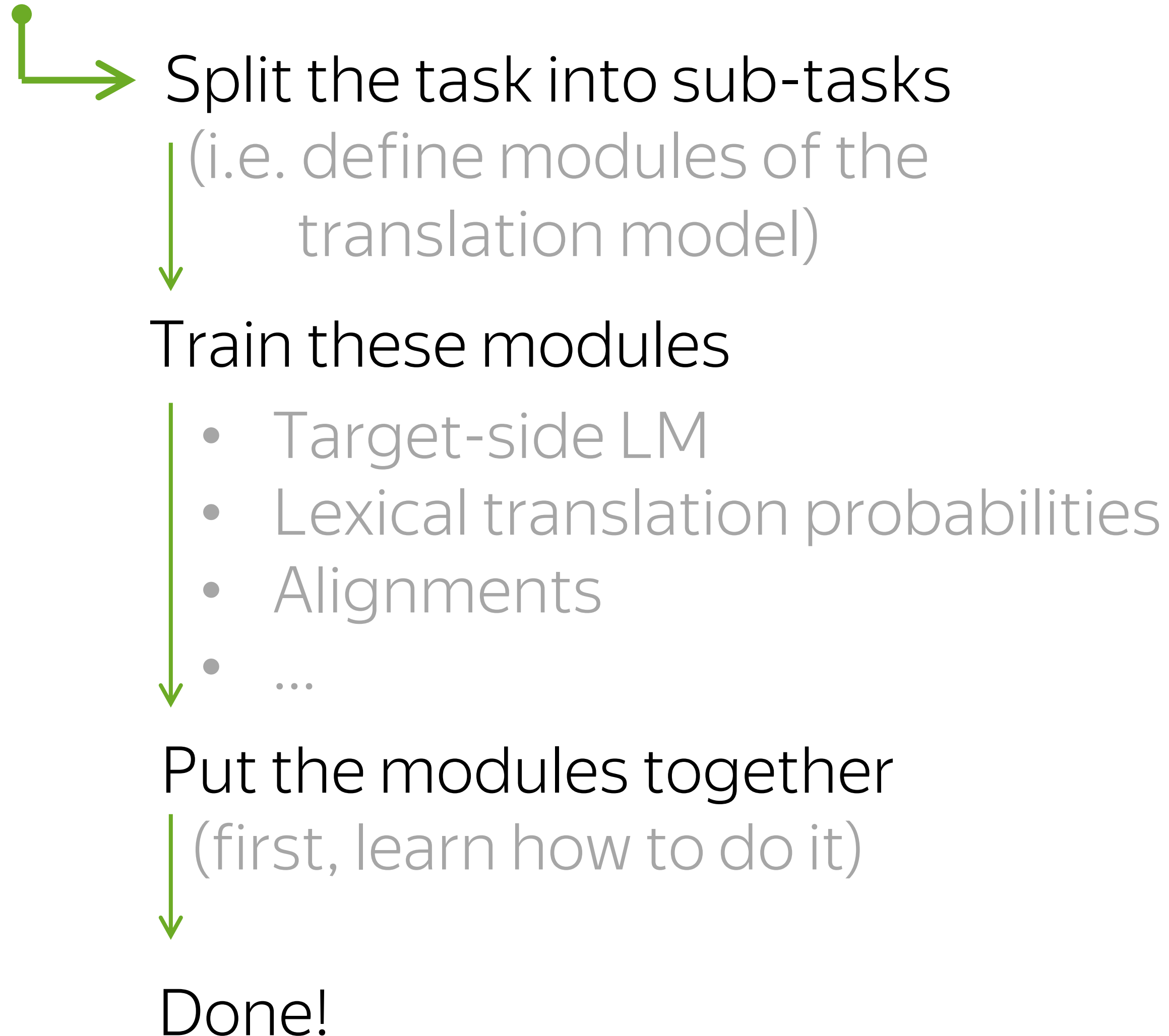
## Statistical MT



## Neural MT

# Take a Step Back: SMT vs NMT

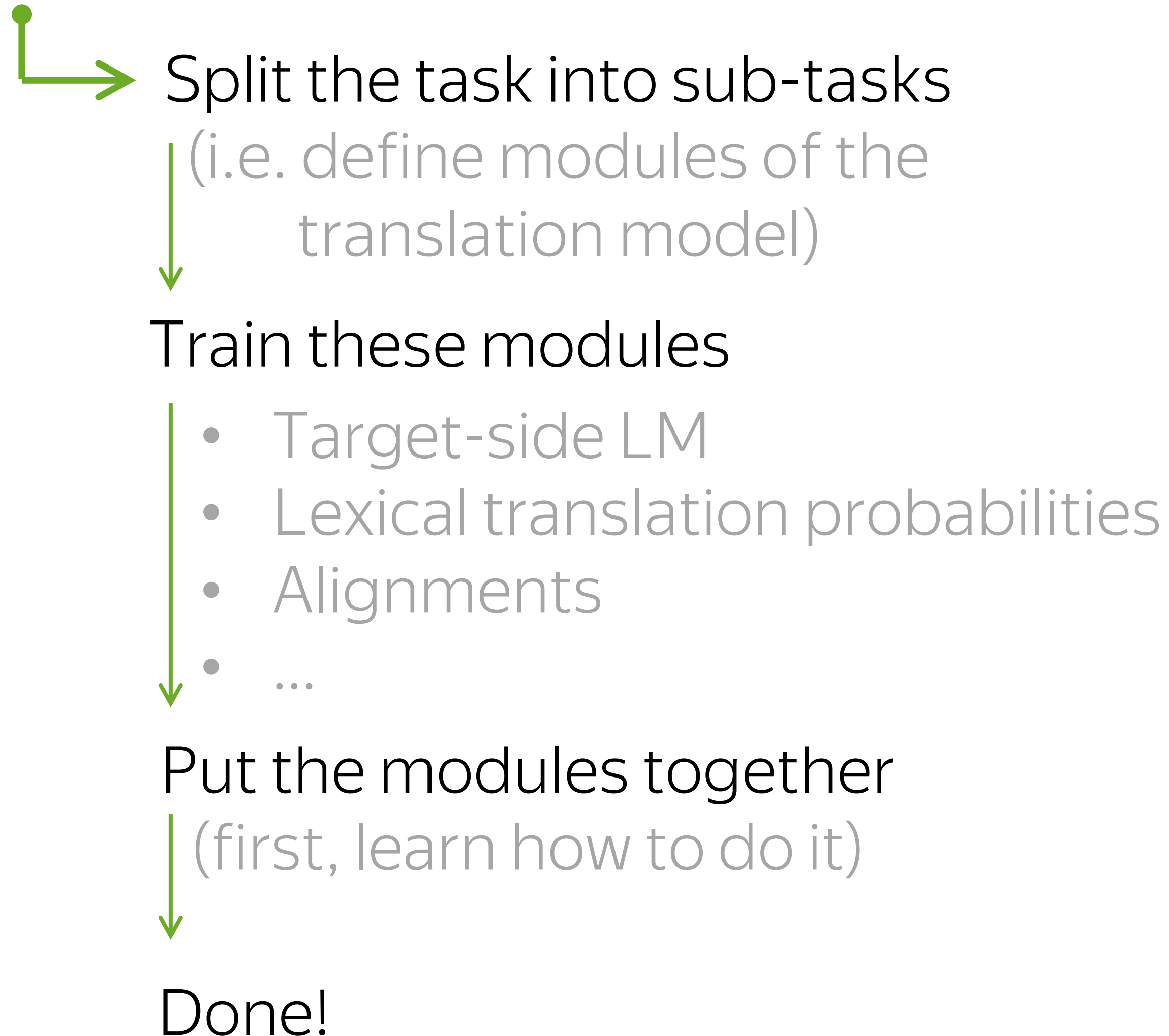
## Statistical MT



## Neural MT

# Take a Step Back: SMT vs NMT

## Statistical MT

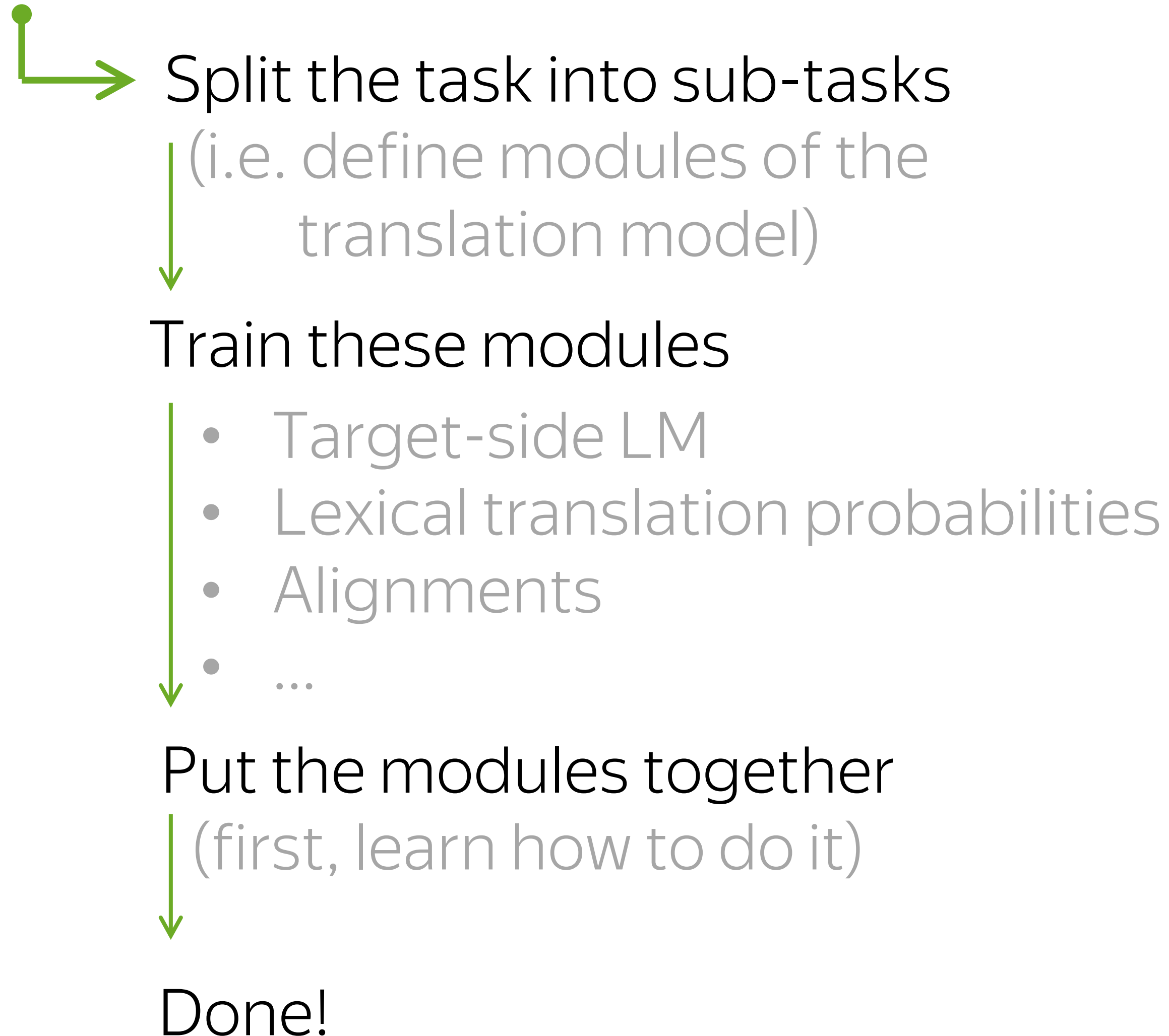


## Neural MT

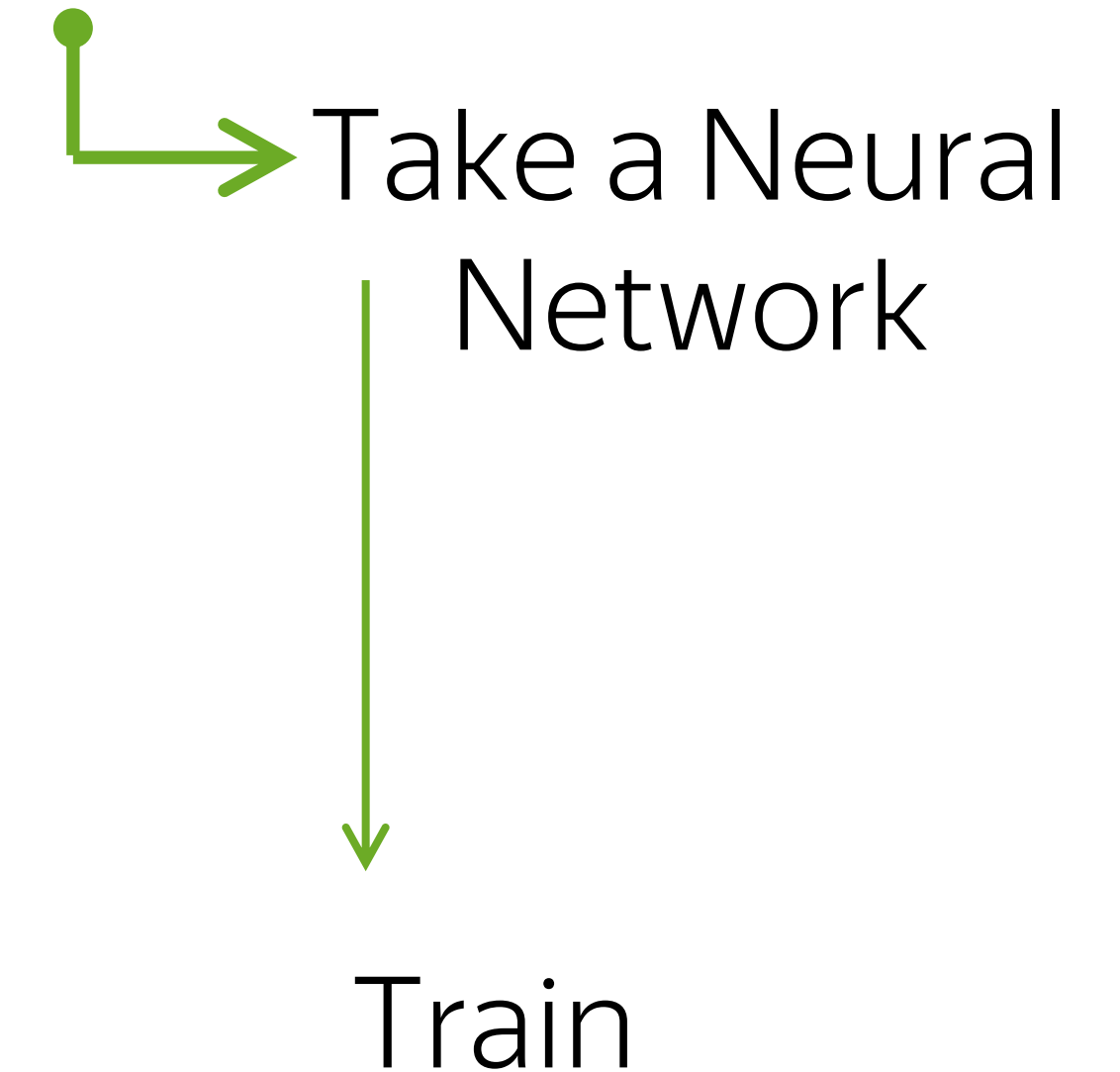


# Take a Step Back: SMT vs NMT

## Statistical MT

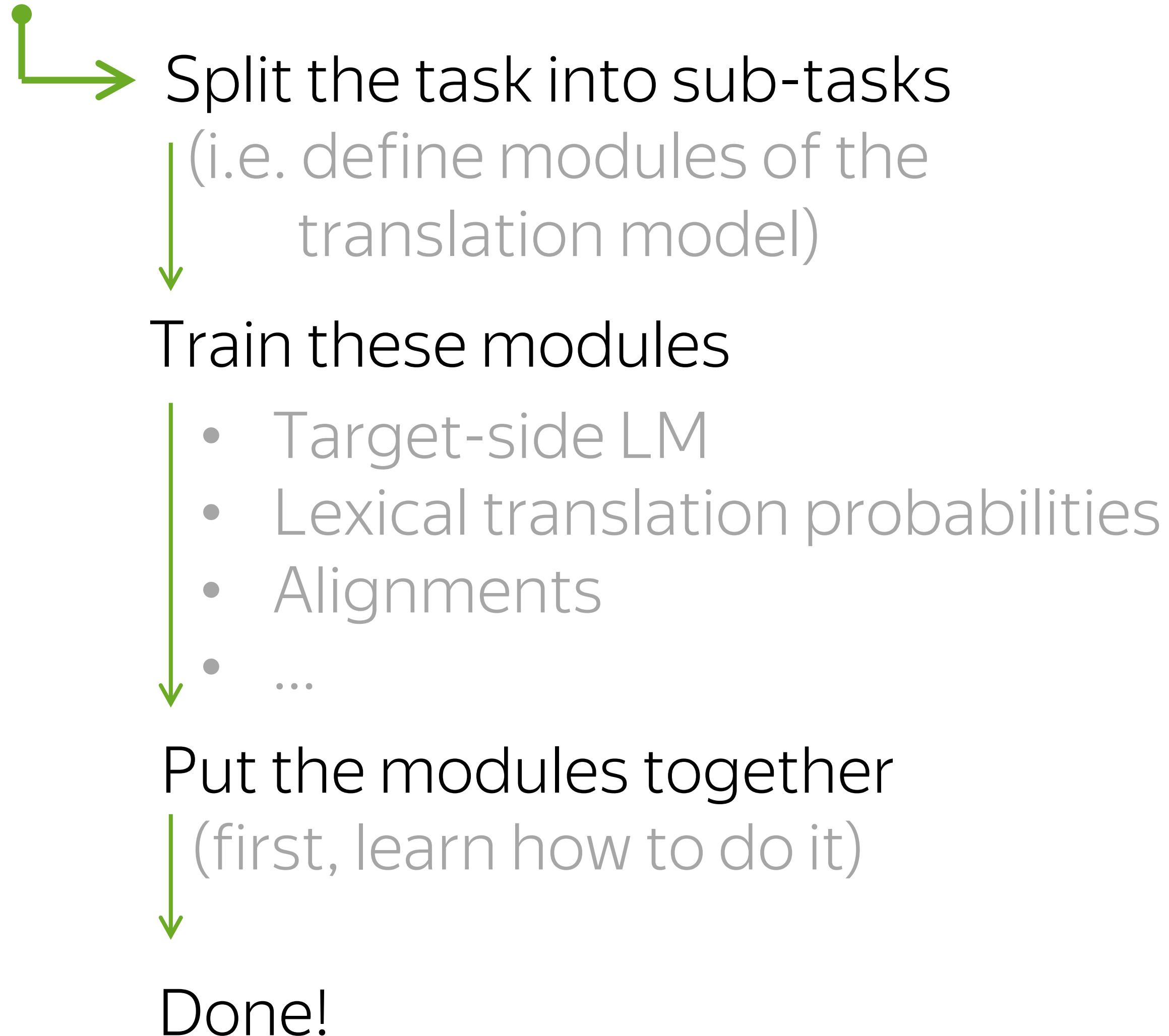


## Neural MT

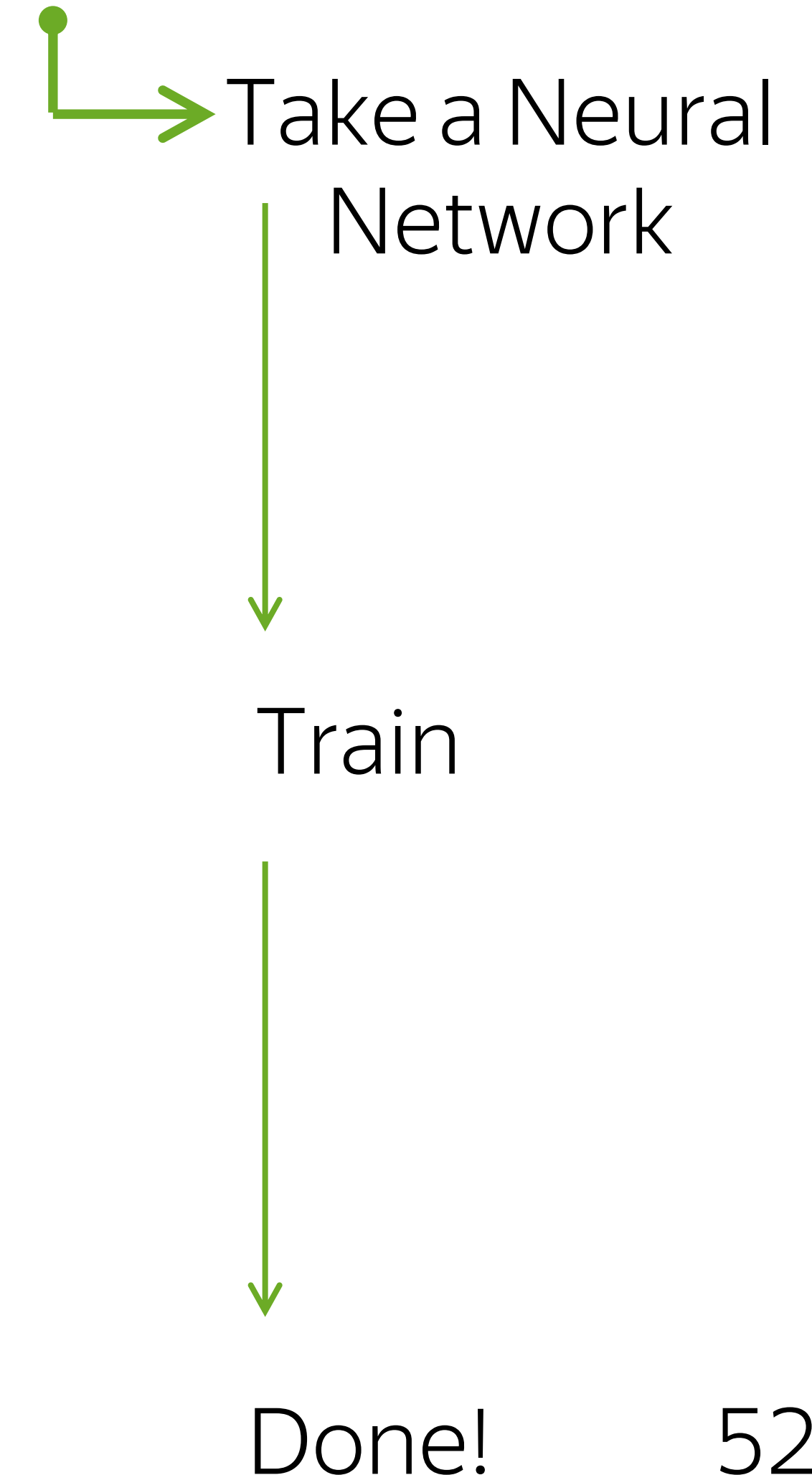


# Take a Step Back: SMT vs NMT

## Statistical MT



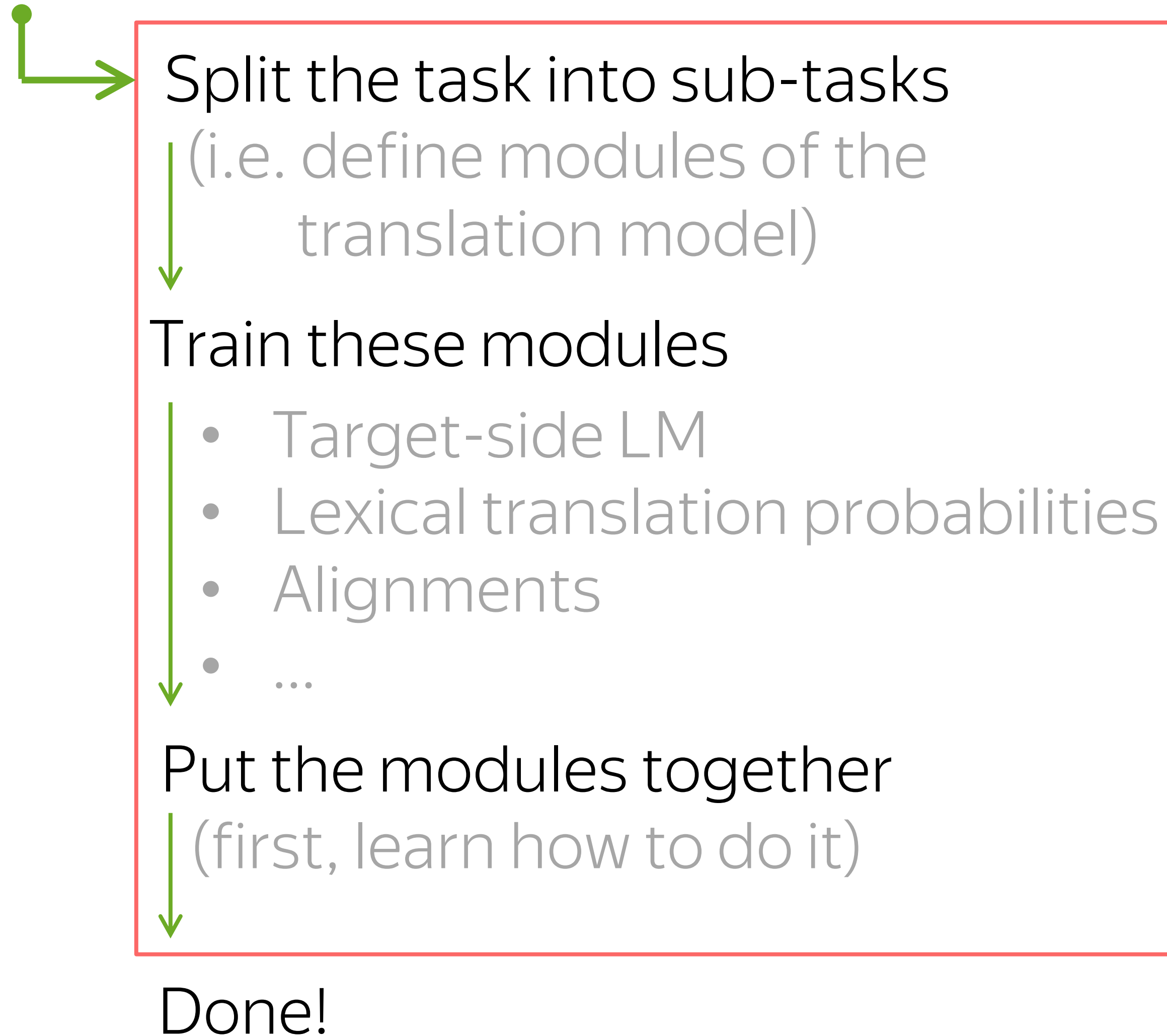
## Neural MT



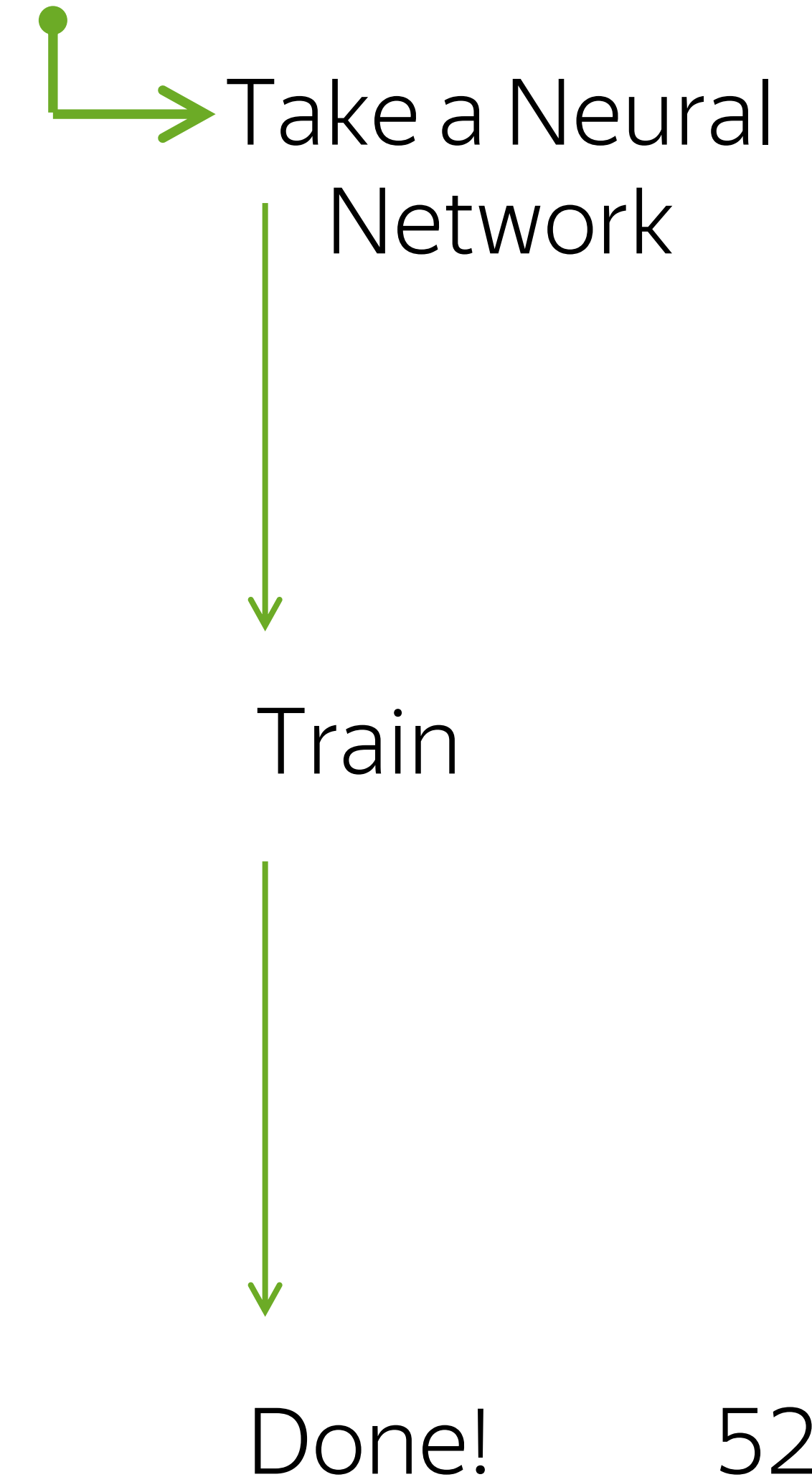


# Take a Step Back: SMT vs NMT

## Statistical MT

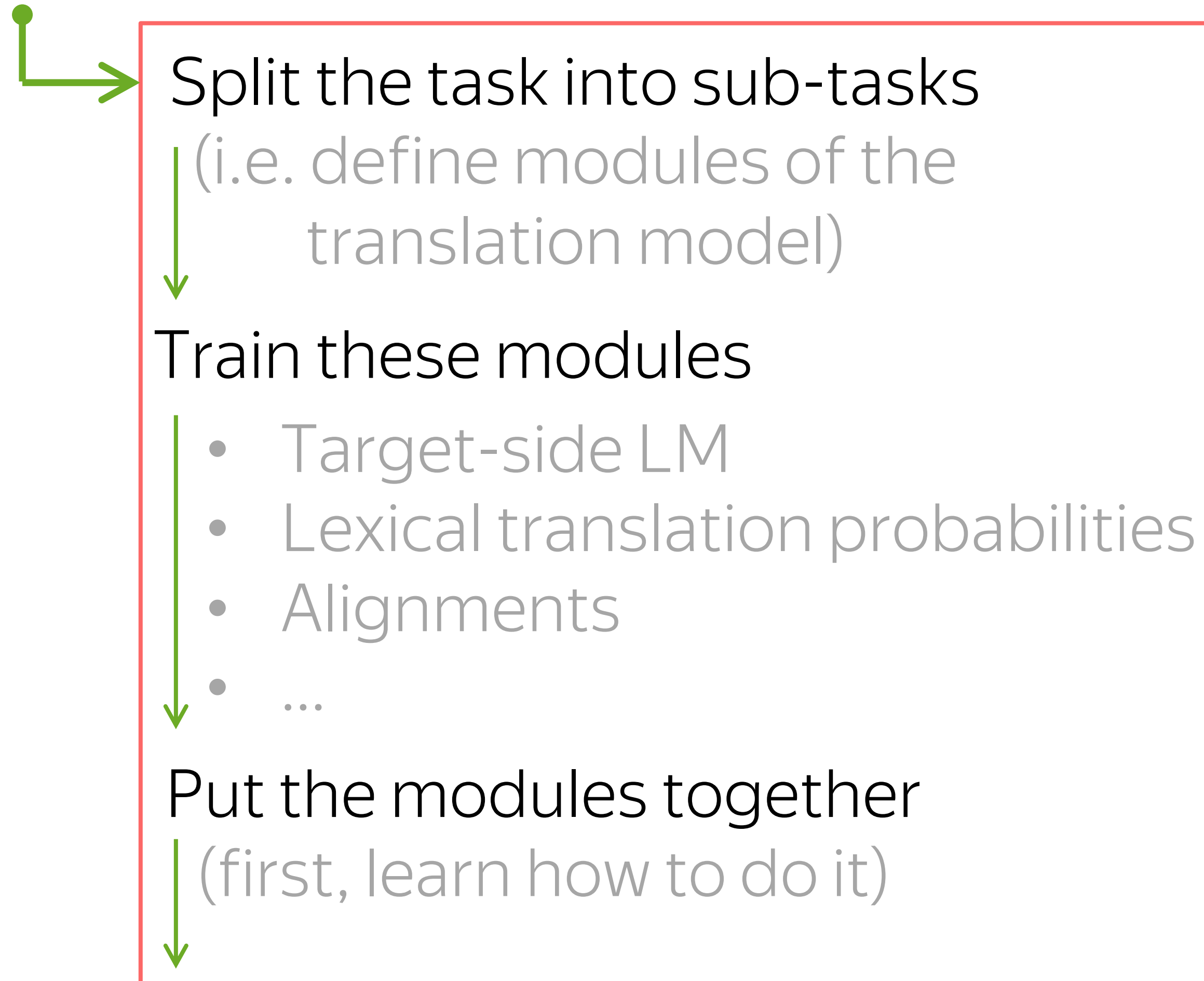


## Neural MT



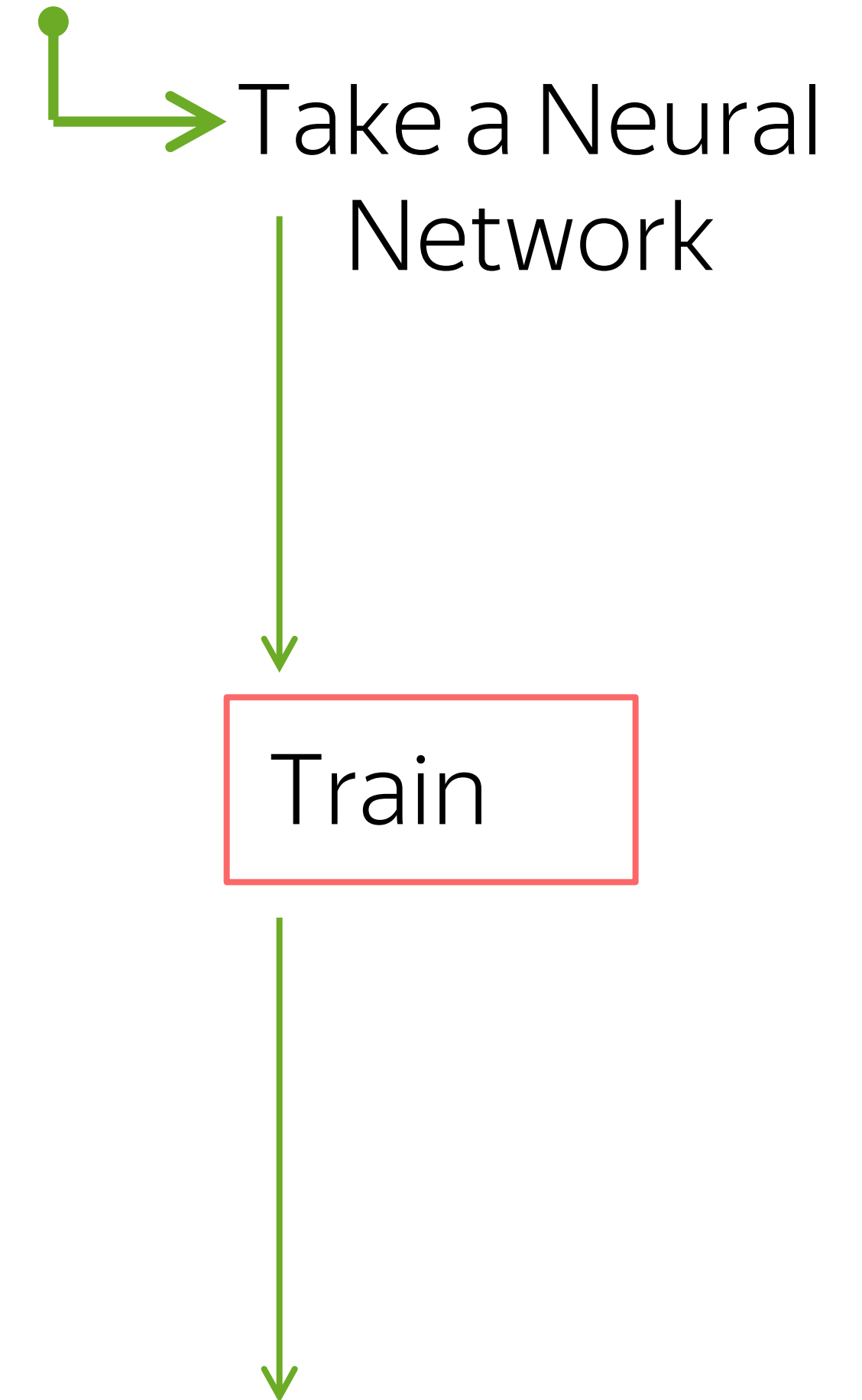
# Take a Step Back: SMT vs NMT

## Statistical MT



Done!

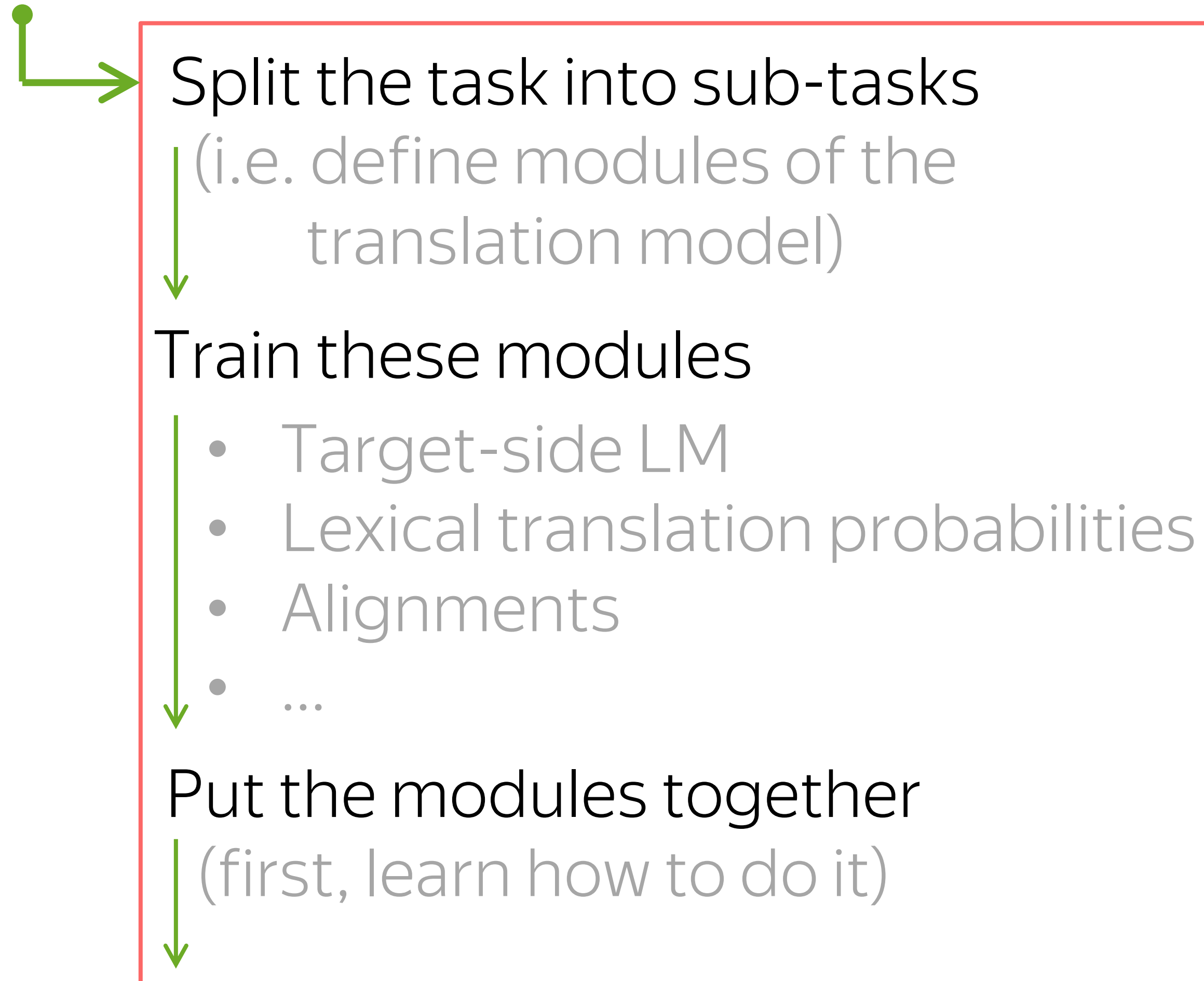
## Neural MT



Done!

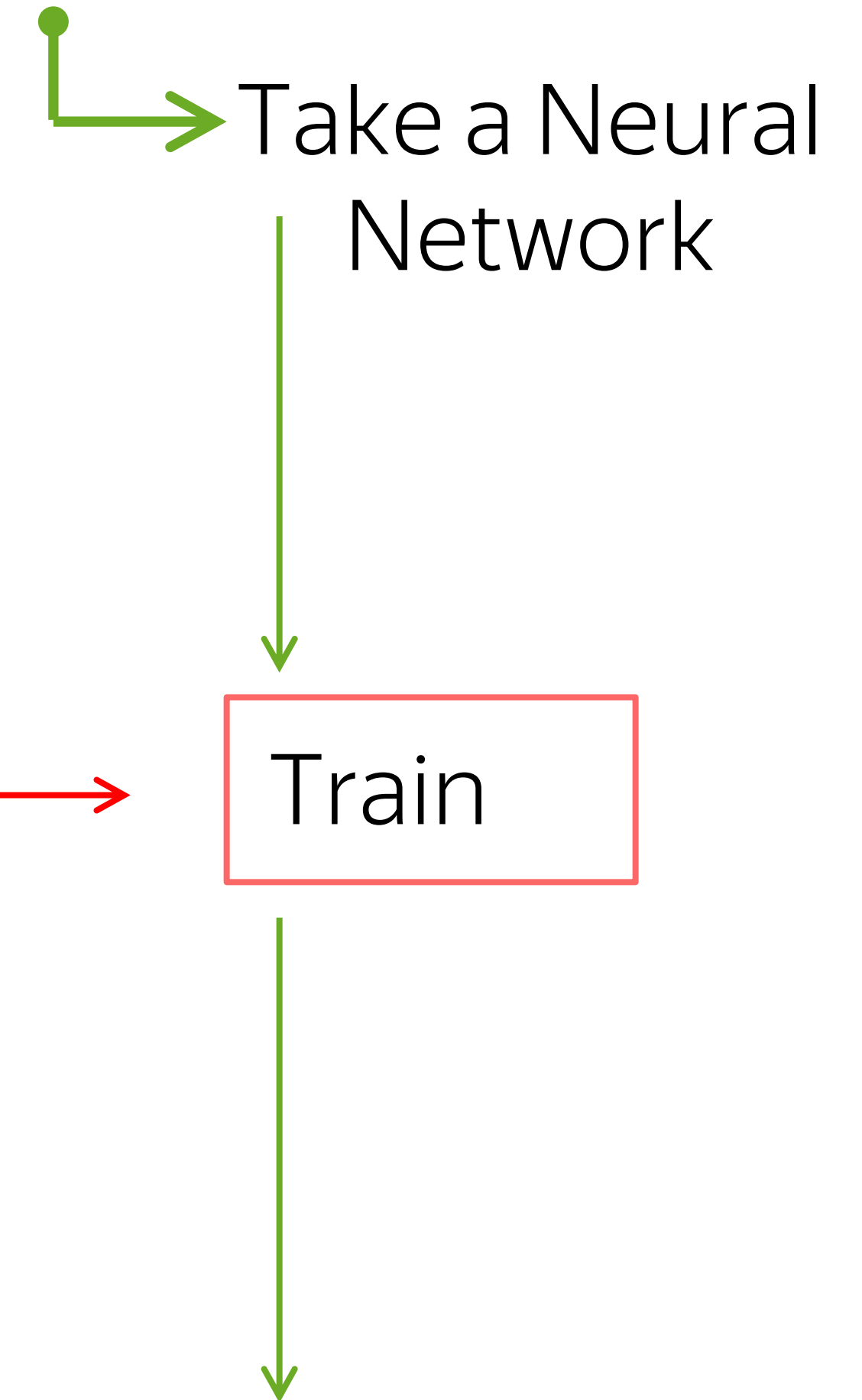
# Take a Step Back: SMT vs NMT

## Statistical MT



Done!

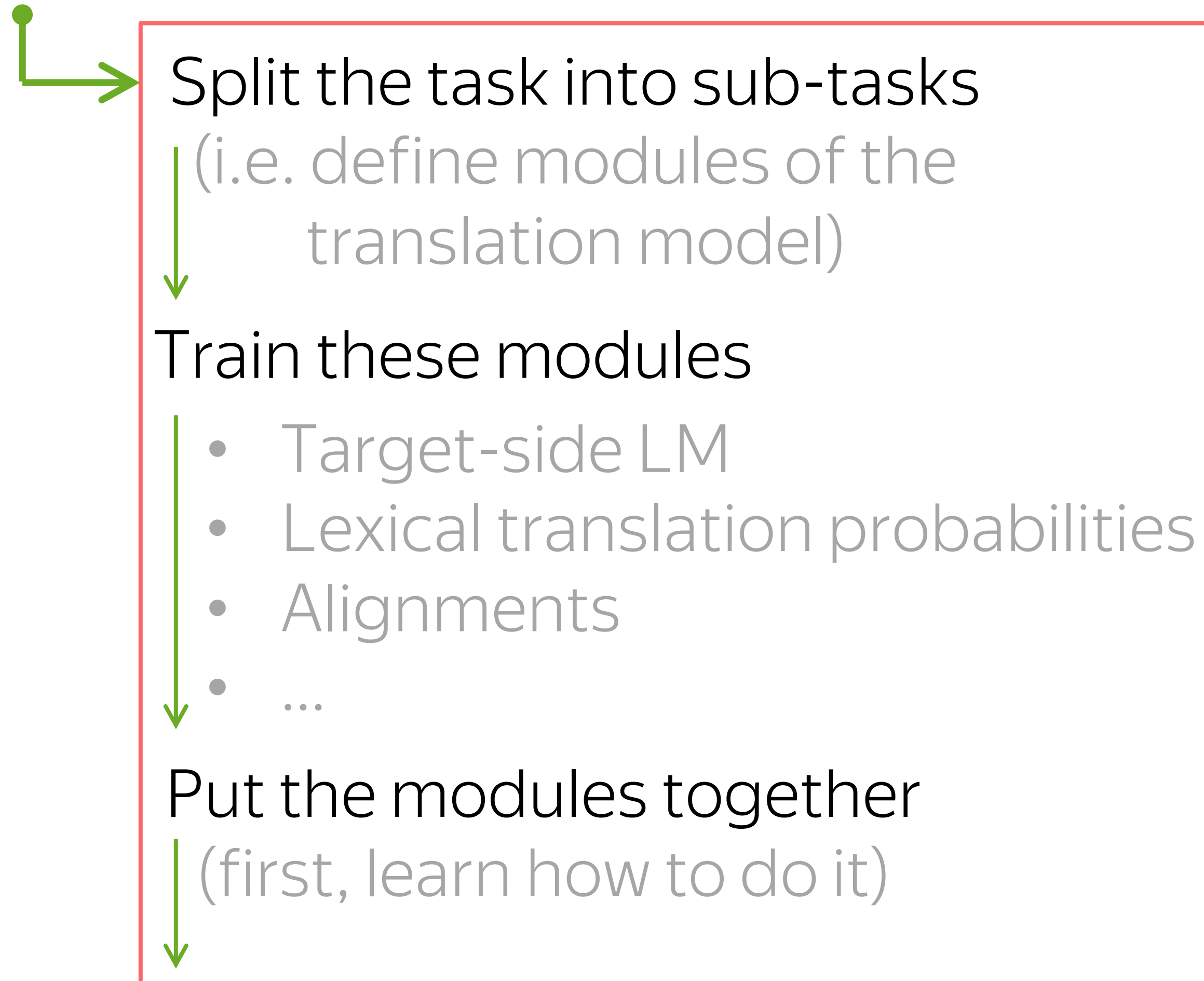
## Neural MT



Done!

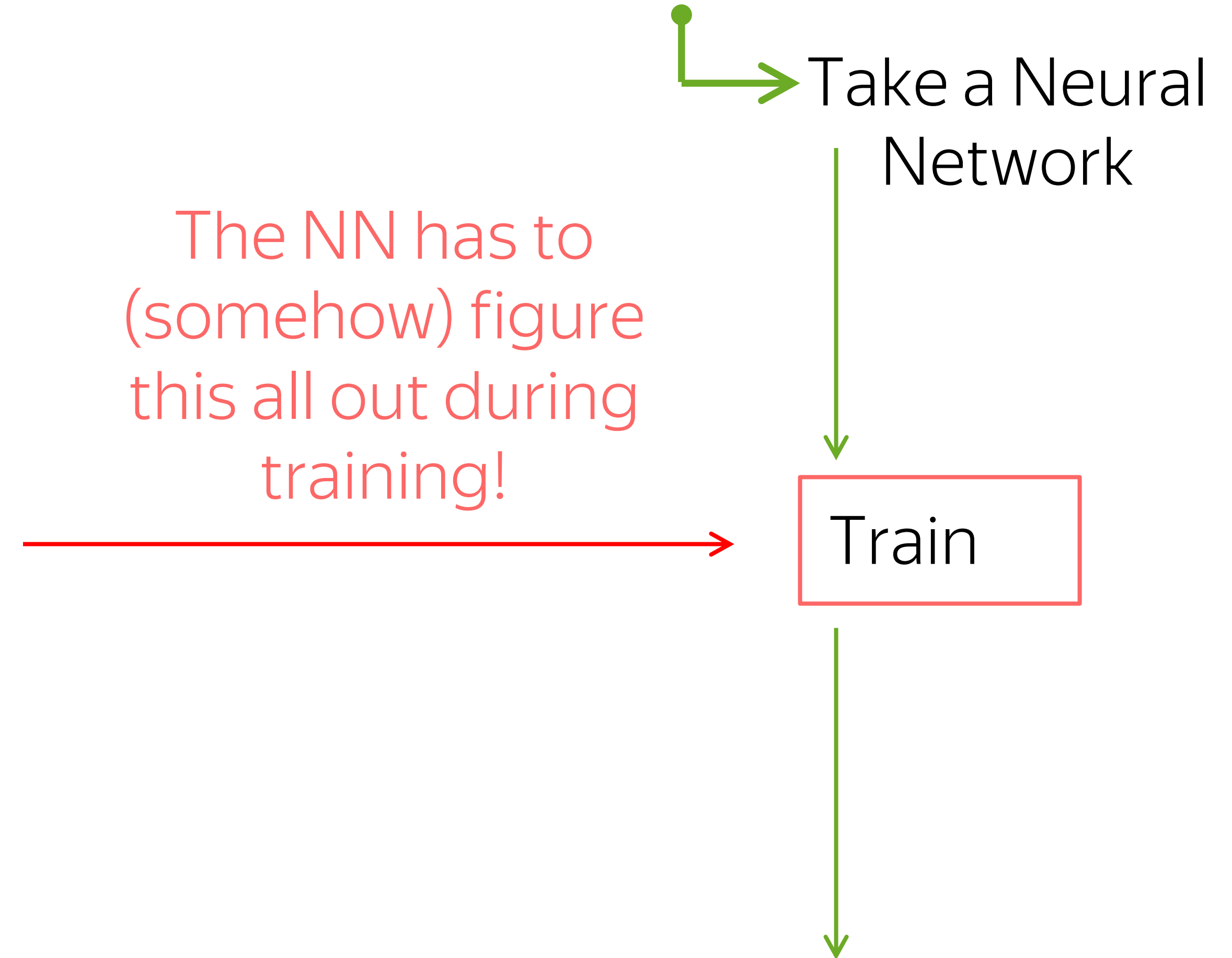
# Take a Step Back: SMT vs NMT

## Statistical MT



Done!

## Neural MT



# NMT Training Process

Neural MT

Take a Neural  
Network

Train

Done!



# NMT Training Process

Neural MT

Take a Neural  
Network

Train

(roughly)

Target LM

Done!

(the order in which NMT learns  
main SMT components)

# NMT Training Process

Neural MT

Take a Neural  
Network

Train

(roughly)

Target LM → Lexical translation pr.

Done!

(the order in which NMT learns  
main SMT components)

# NMT Training Process

Neural MT

Take a Neural  
Network

Train

(roughly)

Target LM → Lexical translation pr. → Alignments

Done!

(the order in which NMT learns  
main SMT components)

# NMT Training Process

Neural MT

Take a Neural Network

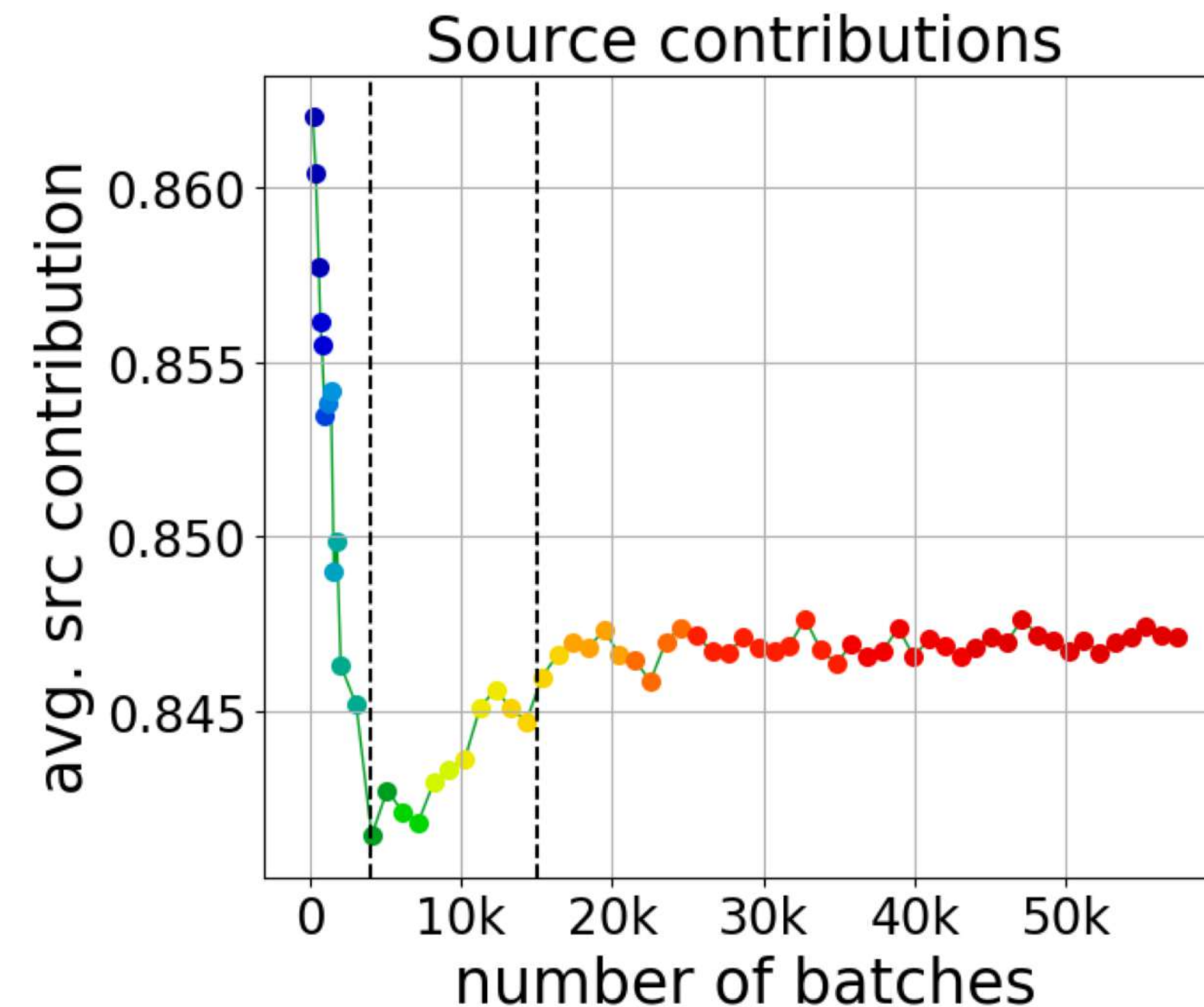
Train

(roughly)

Done!

Target LM → Lexical translation pr. → Alignments

(the order in which NMT learns main SMT components)



(entropy behaves similarly)

# NMT Training Process

Neural MT

Take a Neural Network

Train

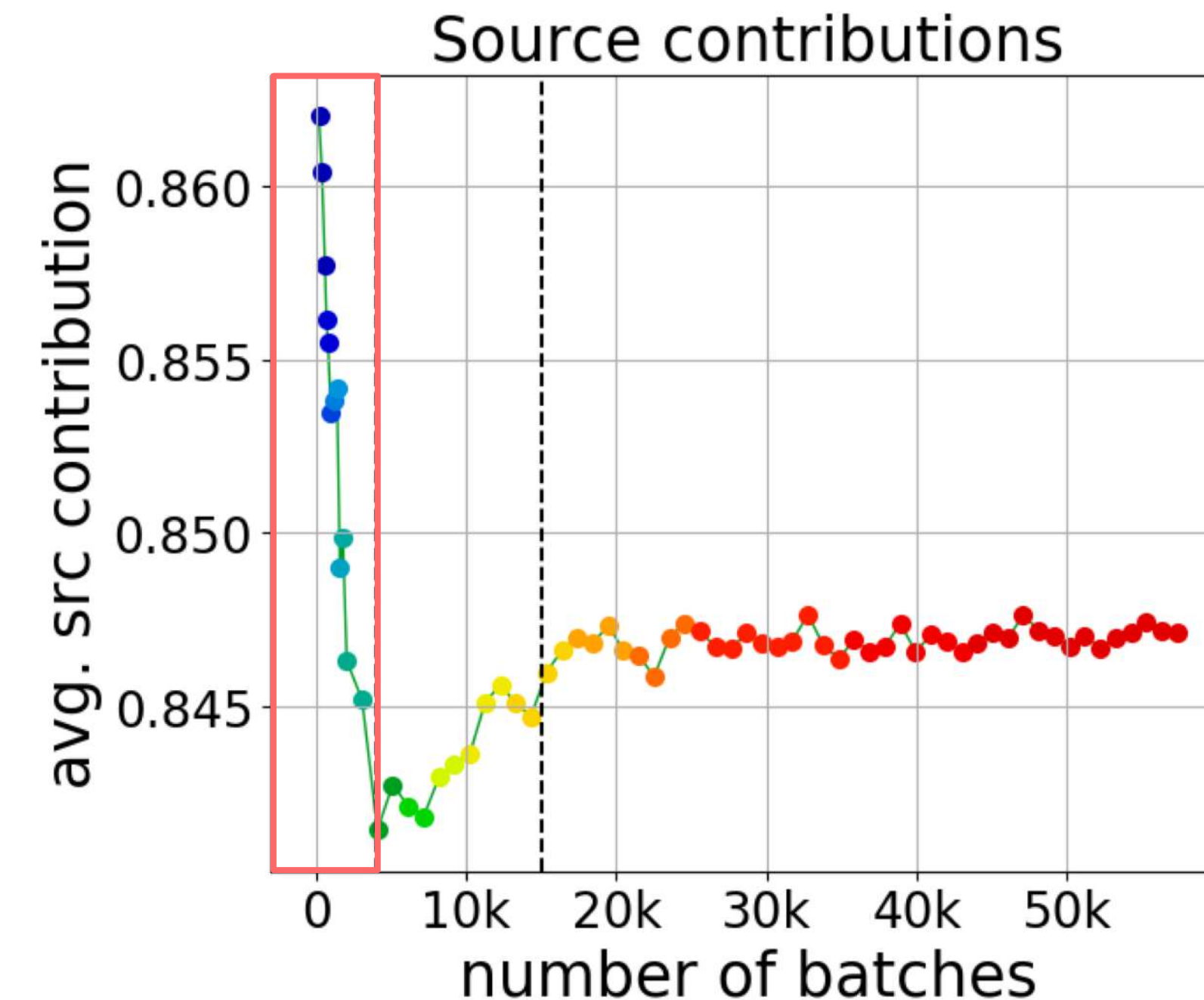
(roughly)

Target LM

Lexical translation pr. → Alignments

Done!

(the order in which NMT learns main SMT components)



(entropy behaves similarly)



# NMT Training Process

Neural MT

Take a Neural Network

Train

(roughly)

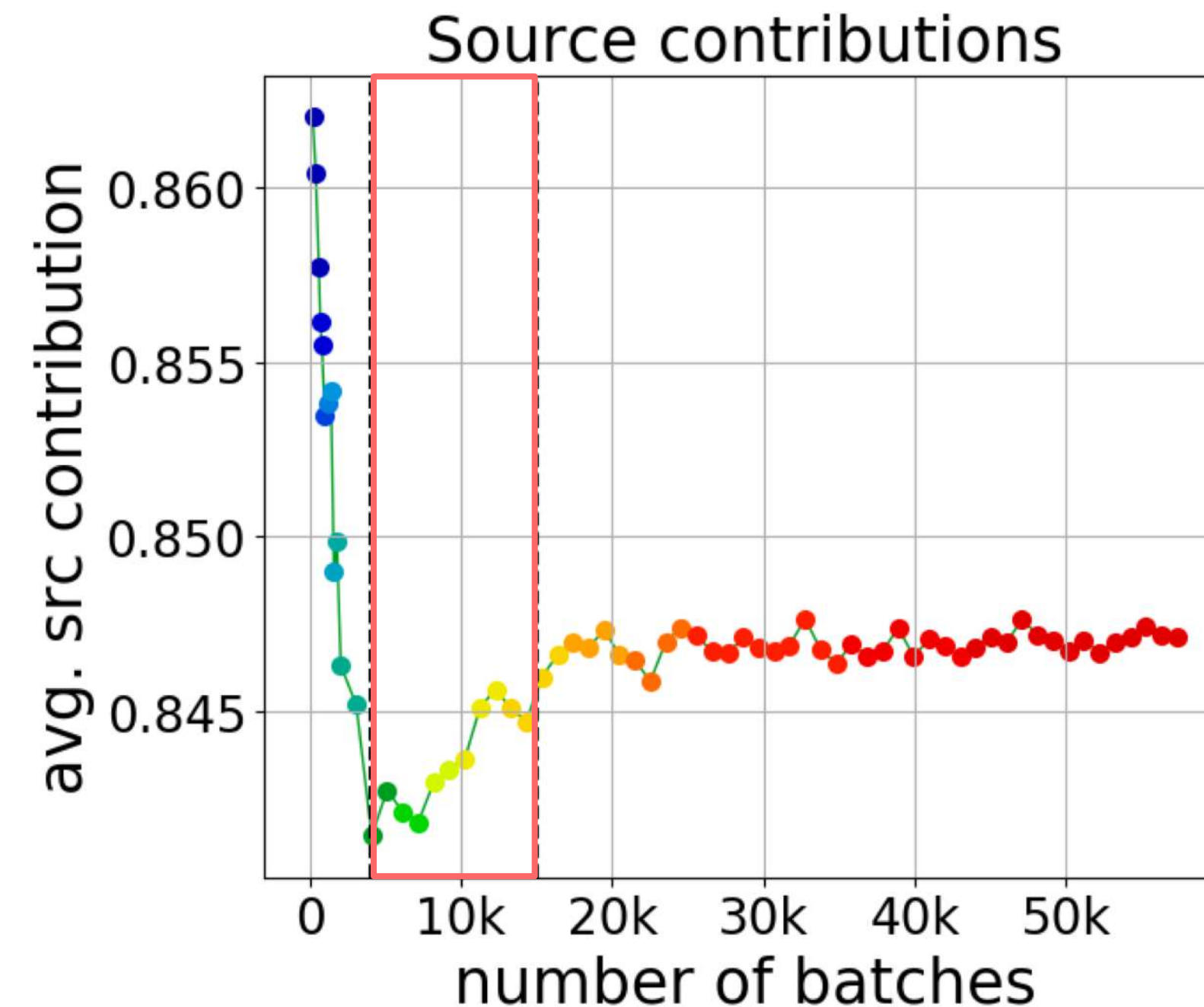
Target LM

Lexical translation pr.

Alignments

Done!

(the order in which NMT learns main SMT components)



(entropy behaves similarly)

# NMT Training Process

Neural MT

Take a Neural Network

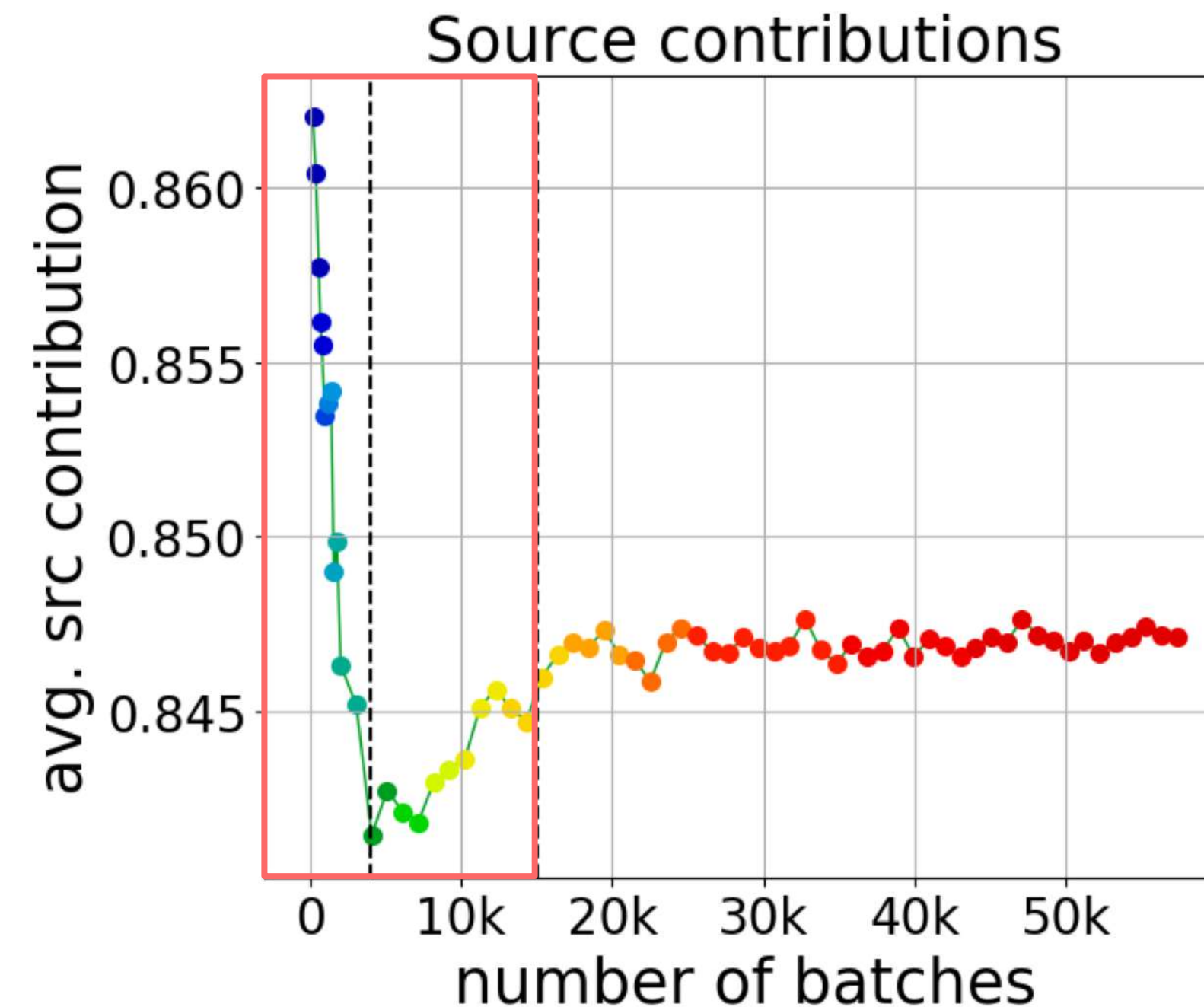
Train

Done!

(roughly)

Target LM → Lexical translation pr. → Alignments

(the order in which NMT learns main SMT components)



(entropy behaves similarly)

(almost) word-by-word translation

# NMT Training Process

Neural MT

Take a Neural Network

Train

(roughly)

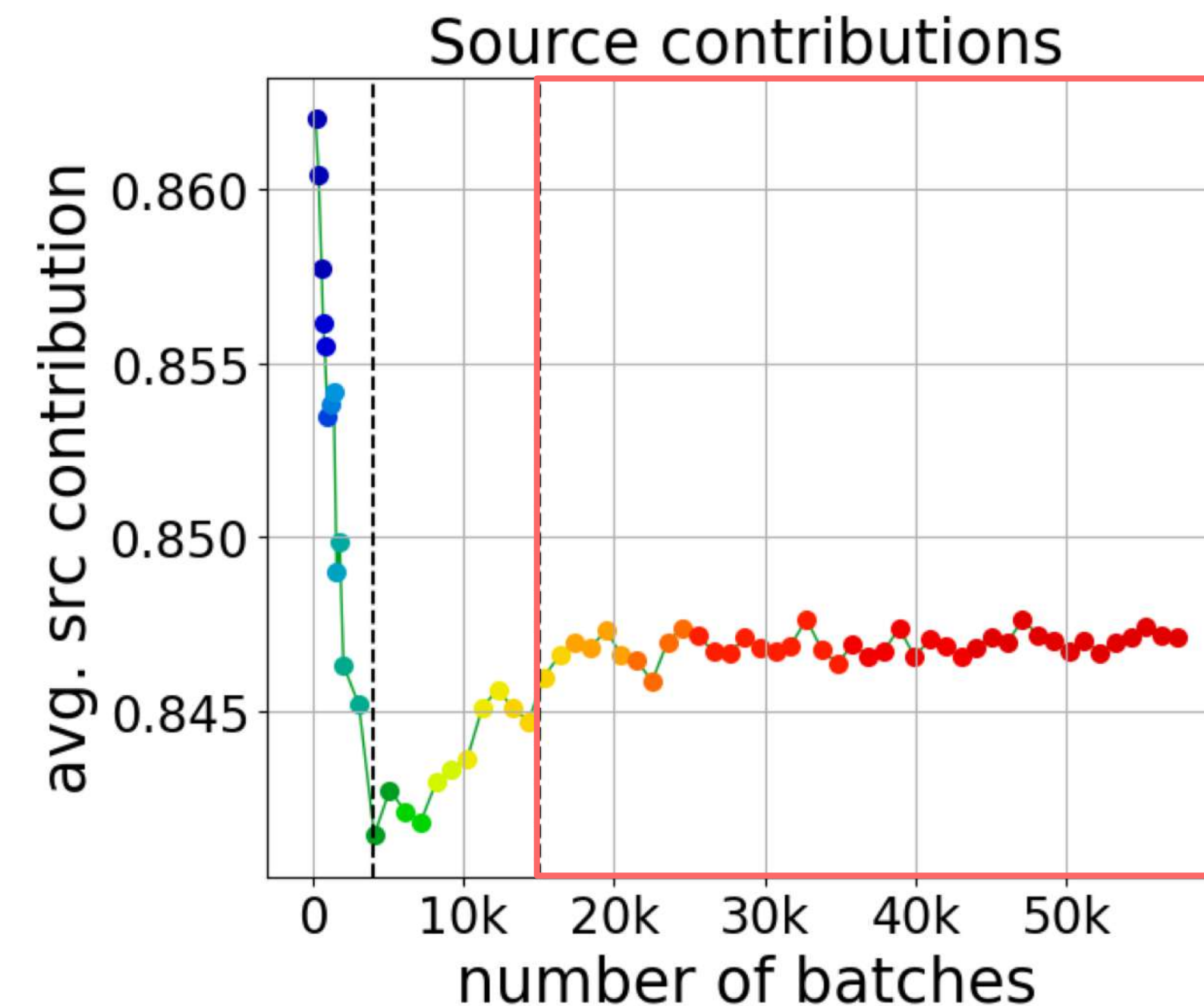
Target LM

Lexical translation pr.

Alignments

Done!

(the order in which NMT learns main SMT components)



(entropy behaves similarly)

# Why is this interesting/useful?

- Sometimes, complexity level (or regularity) of the data is important

(Zhou et al, ICLR 2020, Ren et al, ICLR 2020)

# Why is this interesting/useful?

- Sometimes, complexity level (or regularity) of the data is important

(Zhou et al, ICLR 2020, Ren et al, ICLR 2020)



Translations from specific stages in training may be useful



# Why is this interesting/useful?

- Sometimes, complexity level (or regularity) of the data is important

([Zhou et al, ICLR 2020](#), [Ren et al, ICLR 2020](#))



Translations from specific stages in training may be useful

- SMT-inspired model modifications often help

(using target LM/lexical tables/alignments, modeling phrases, etc)

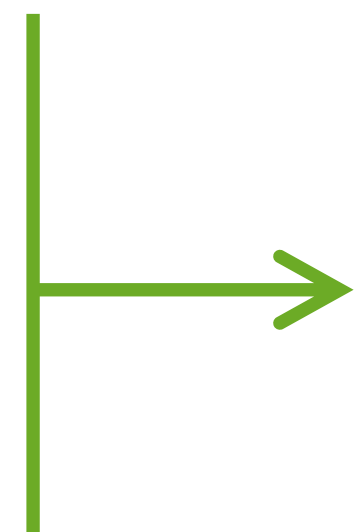
# Why is this interesting/useful?

- Sometimes, complexity level (or regularity) of the data is important  
([Zhou et al, ICLR 2020](#), [Ren et al, ICLR 2020](#))



Translations from specific stages in training may be useful

- SMT-inspired model modifications often help  
(using target LM/lexical tables/alignments, modeling phrases, etc)



The analysis can help for (i) understanding the NMT model, and/or (ii) modeling

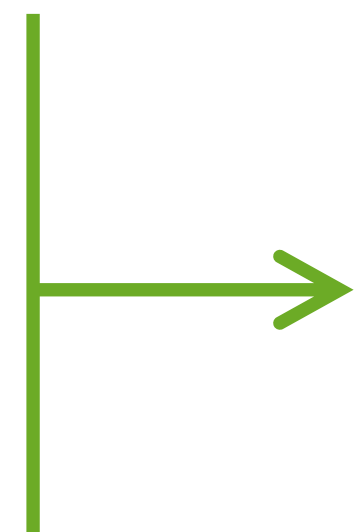
# Why is this interesting/useful?

- Sometimes, complexity level (or regularity) of the data is important  
([Zhou et al, ICLR 2020](#), [Ren et al, ICLR 2020](#))



Translations from specific stages in training may be useful

- SMT-inspired model modifications often help  
(using target LM/lexical tables/alignments, modeling phrases, etc)



The analysis can help for (i) understanding the NMT model, and/or (ii) modeling

- Your options?

# Conclusions



# Conclusions

- We show that LRP can be used to evaluate relative source and target contributions to NMT predictions
- Some of the findings are:
  - with more data, models use source more and are more confident in the choice of important tokens
  - models suffering from exposure bias are more prone to over-relying on target history
  - training process is not monotonic with several distinct stages



# Conclusions

- We show that LRP can be used to evaluate relative source and target contributions to NMT predictions
  - Some of the findings are:
    - with more data, models use source more and are more confident in the choice of important tokens
    - models suffering from exposure bias are more prone to over-relying on target history
    - training process is not monotonic with several distinct stages
- Target LM -> Lexical stuff -> alignments  
(work in progress)

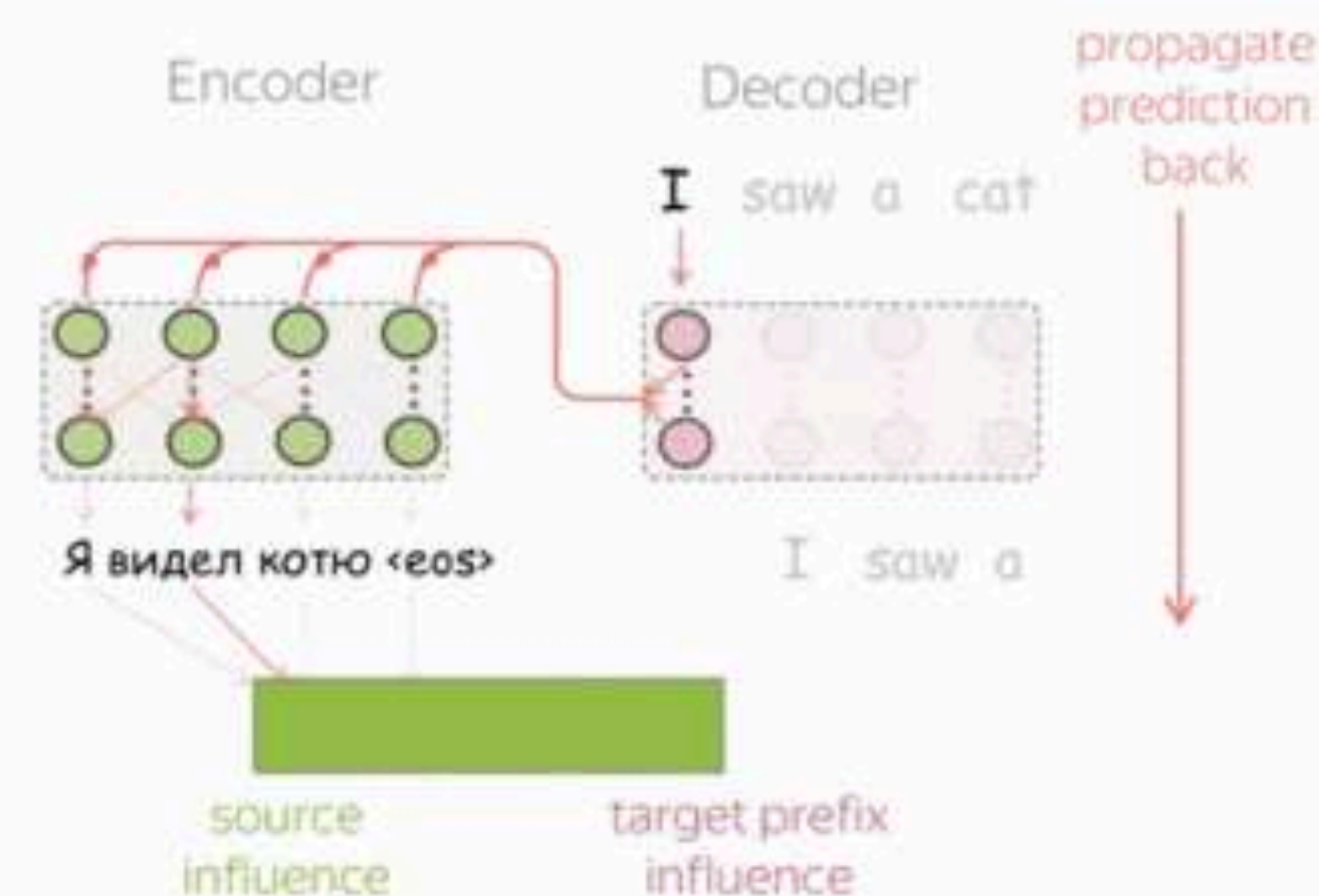
# Source and Target Contributions to NMT Predictions

## Source and Target Contributions to NMT Predictions

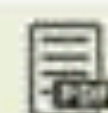
This is a post for the paper [Analyzing the Source and Target Contributions to Predictions in Neural Machine Translation](#).

In NMT, the generation of a target token is based on two types of context: the source and the prefix of the target sentence. We show how to evaluate the relative contributions of source and target to NMT predictions and find that:

- models suffering from exposure bias are more prone to over-relying on target history (and hence to hallucinating) than the ones where the exposure bias is mitigated;
- models trained with more data rely on the source more and do it more confidently;
- the training process is non-monotonic with several distinct stages.



→ read more



read paper

</> view code

October 2020

What else?



# Context-Aware NMT

- ACL 2018: Context-Aware NMT Learns Anaphora Resolution

When interaction with context is limited, what does a model learn?

# Context-Aware NMT

- ACL 2018: Context-Aware NMT Learns Anaphora Resolution

When interaction with context is limited, what does a model learn?

- ACL 2019: When a Good Translation is Wrong in Context

Which phenomena are the most important and how to evaluate them?

Usually, you have lots of sentence-level parallel data and only a bit of document-level. What can you do?



# Context-Aware NMT

- ACL 2018: Context-Aware NMT Learns Anaphora Resolution

When interaction with context is limited, what does a model learn?

- ACL 2019: When a Good Translation is Wrong in Context

Which phenomena are the most important and how to evaluate them?

Usually, you have lots of sentence-level parallel data and only a bit of document-level. What can you do?

- EMNLP 2019: Context-Aware Monolingual Repair for NMT

Context-Aware NMT model without parallel document-level data

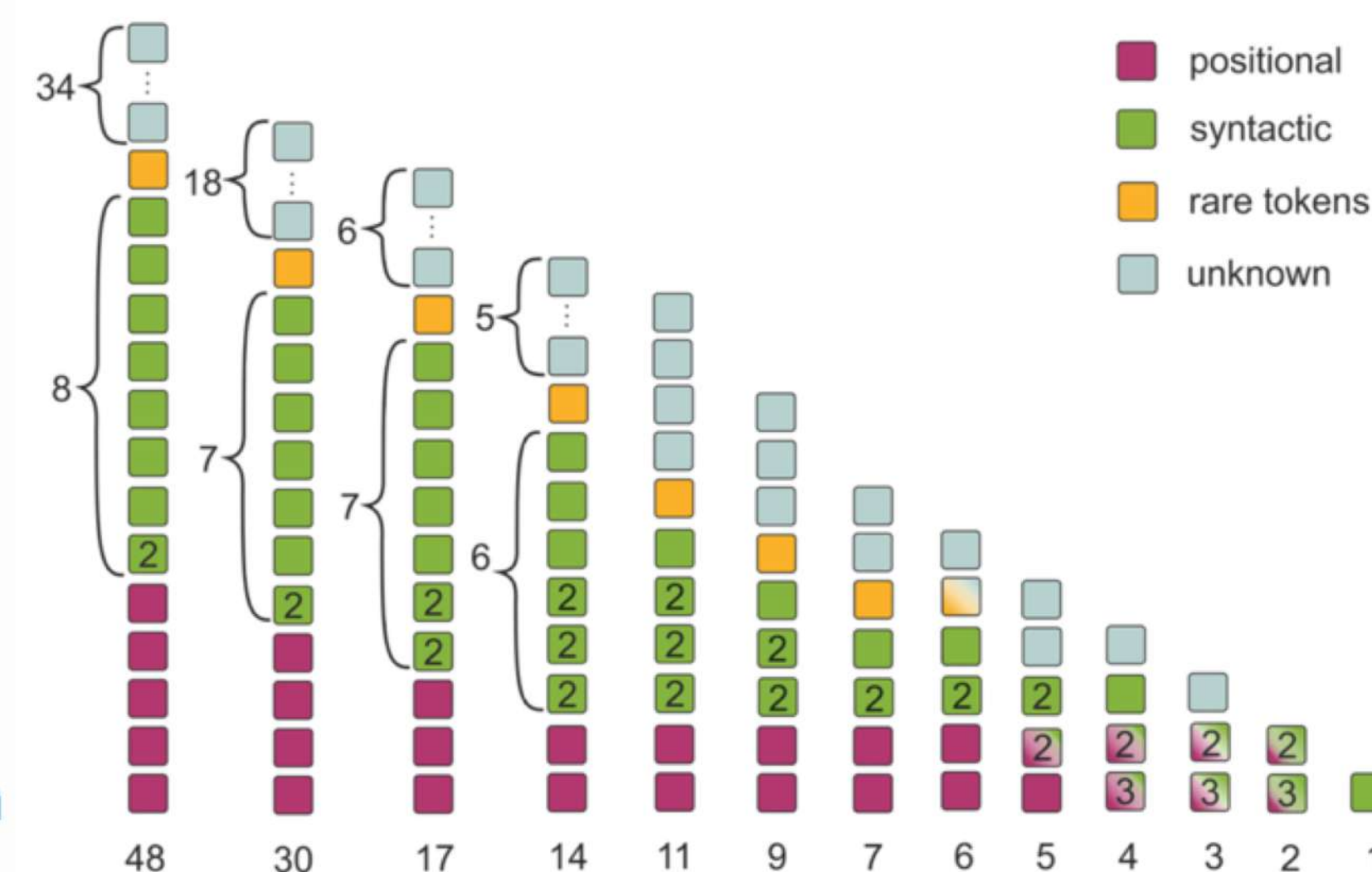
# (ACL 2019) Transformer's Attention Heads: Important are Interpretable, the Rest can be Pruned

## The Story of Heads

This is a post for the ACL 2019 paper [Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned](#).

From this post, you will learn:

- how we evaluate the importance of attention heads in Transformer
- which functions the most important encoder heads perform
- how we prune the vast majority of attention heads in Transformer without seriously affecting quality
- which types of model attention are most sensitive to the number of attention heads and on which layers



→ read more



read paper

</> view code

June 2019



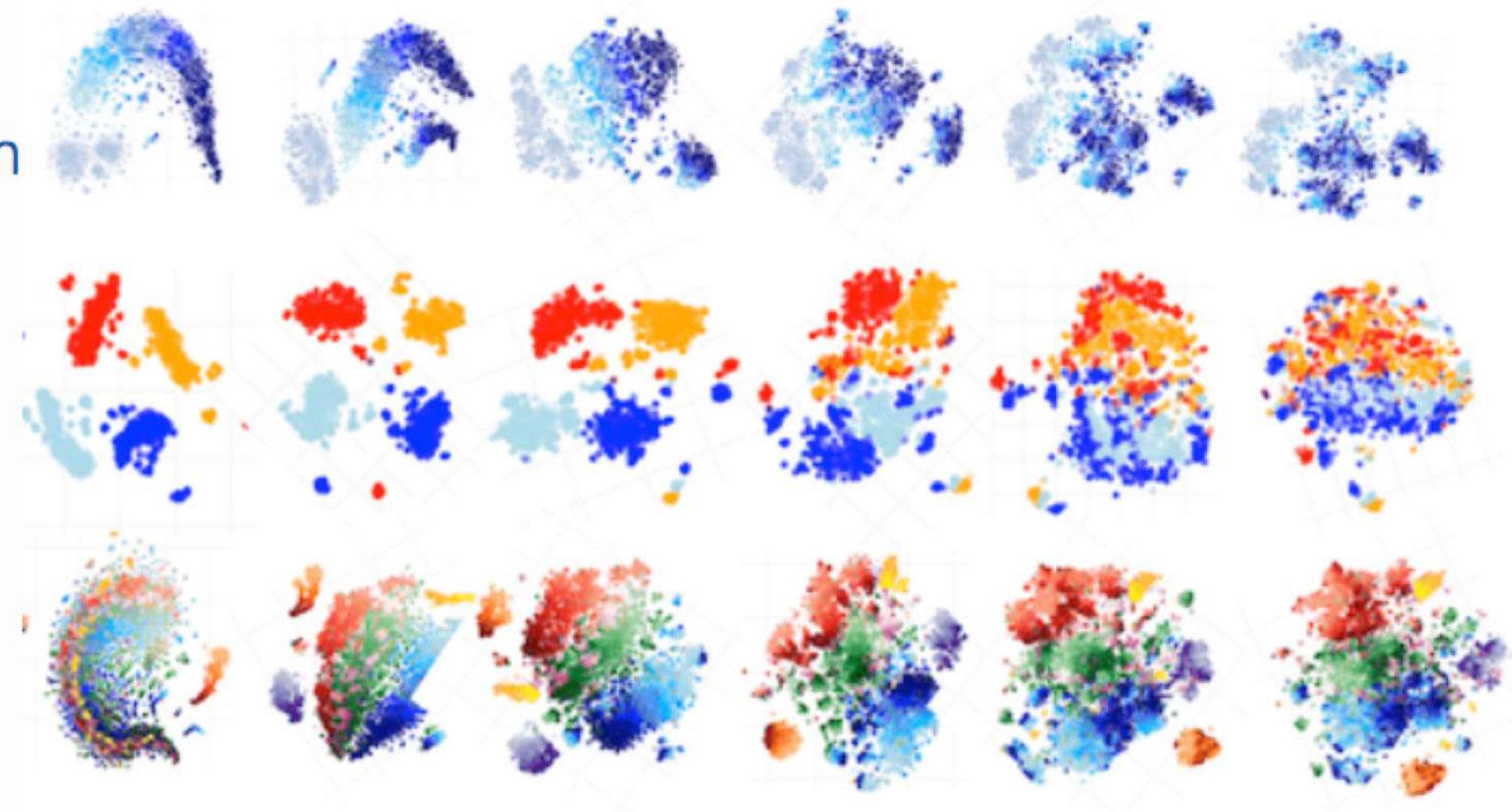
# (EMNLP 2019) Evolution of Representations in the Transformer

## Evolution of Representations in the Transformer

This is a post for the EMNLP 2019 paper [The Bottom-up Evolution of Representations in the Transformer: A Study with Machine Translation and Language Modeling Objectives](#).

We look at the evolution of representations of individual tokens in Transformers trained with different training objectives (MT, LM, MLM - BERT-style) from the [Information Bottleneck](#) perspective and show, that:

- LMs gradually forget past when forming predictions about future;
- for MLMs, the evolution proceeds in two stages of **context encoding** and **token reconstruction**;
- MT representations get refined with context, but less processing is happening.



→☰ read more

📄 read paper

September 2019



# (EMNLP 2020) Information-Theoretic Probing with MDL

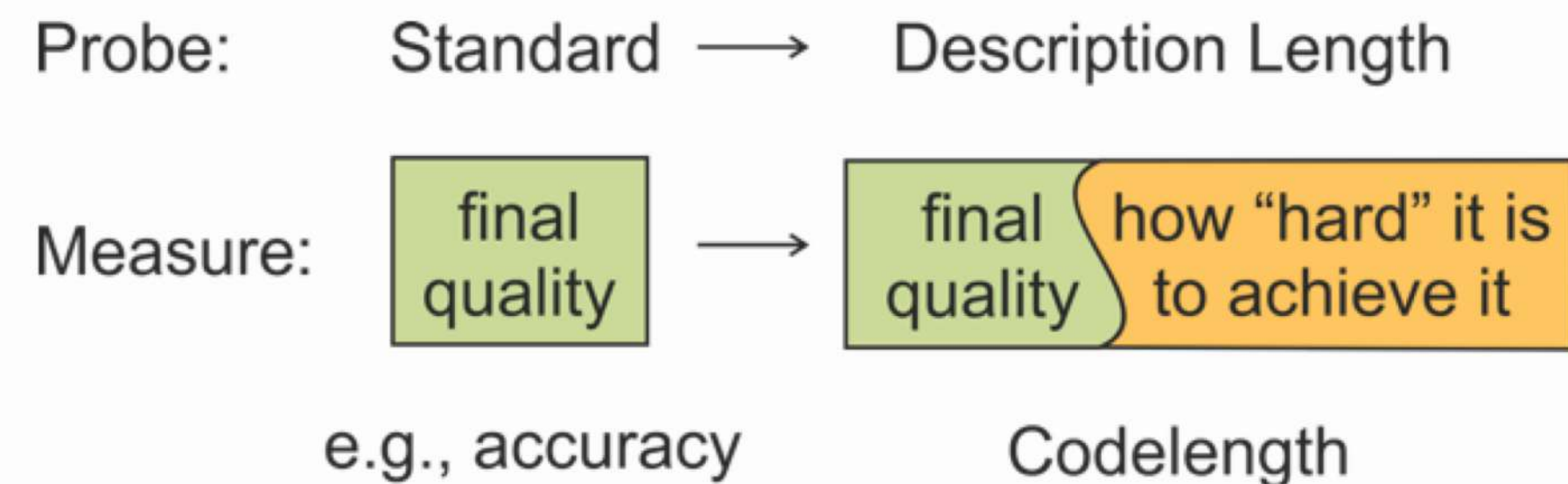
## Information-Theoretic Probing with MDL

This is a post for the EMNLP 2020 paper [Information-Theoretic Probing with Minimum Description Length](#).

Probing classifiers often fail to adequately reflect differences in representations and can show different results depending on hyperparameters.

As an alternative to the standard probes,

- we propose information-theoretic probing which measures **minimum description length** (MDL) of labels given representations;
- we show that MDL characterizes both **probe quality** and the **amount of effort** needed to achieve it;
- we explain how to easily measure MDL on top of standard probe-training pipelines;
- we show that results of MDL probes are more informative and stable than those of standard probes.



→≡ read more



read paper

</> view code

March 2020



# Students/interns: BPE-Dropout (ACL 2020)

## BPE-Dropout: Simple and Effective Subword Regularization

Ivan Provilkov\*, Dmitrii Emelianenko\*, Elena Voita

u-n-r-e-l-a-t-e-d  
u-n re-l-a-t-e-d  
u-n re-l-at-e-d  
u-n re-l-at-ed  
un re-l-at-ed  
un re-l-ated  
un rel-ated  
un-related  
unrelated

u-n\_r-e-l-a\_t-e\_d  
u-n re-l\_a-t-e\_d  
u-n re\_l-at-e\_d  
un re-l-at-e-d  
un re\_l-at-ed  
un re-lat-ed  
un re-lat-ed  
un relat\_ed

u-n-r-e-l-a\_t-e-d  
u\_n re\_l-a-t-e-d  
u\_n re-l-at-e-d  
u\_n re-l-ate\_d  
u\_n rel-ate\_d  
u\_n relate\_d

u-n\_r\_e\_l-a-t-e-d  
u-n-r\_e-l-at-e-d  
u-n-r\_e-l\_at\_ed  
un-r-e-l-at-ed  
un re-l\_at-ed  
un re-l-ated  
un rel\_ated

BPE: deterministic

BPE-Dropout: stochastic  
(use in training, get profit)

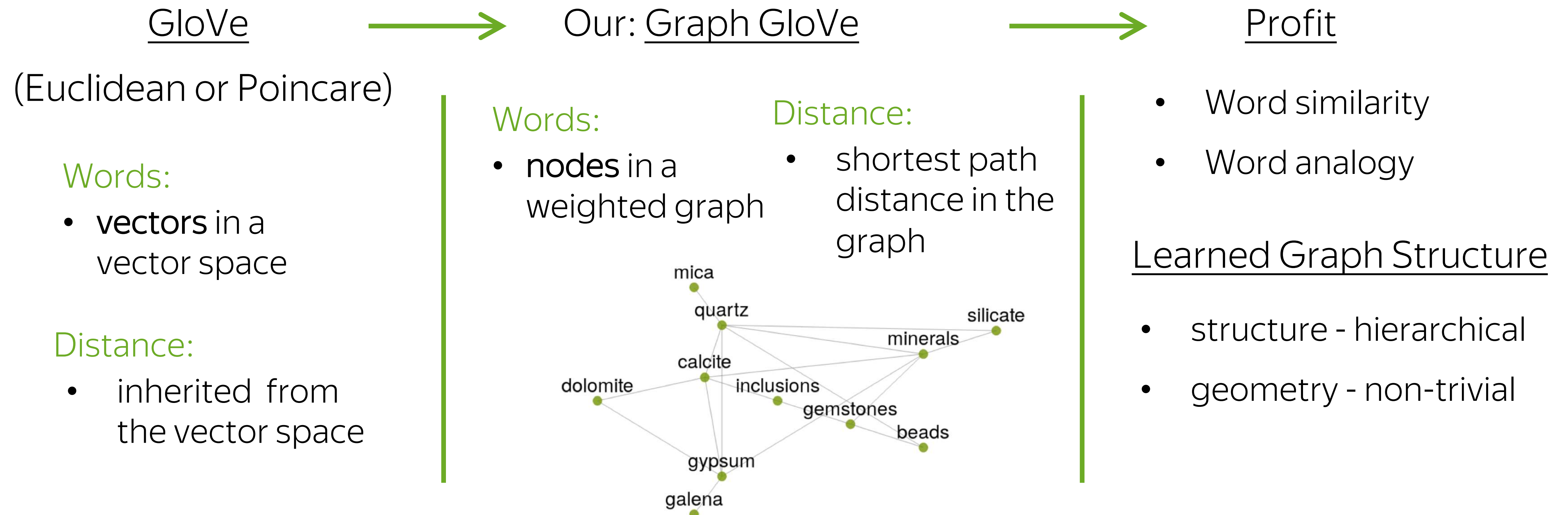
\*equal contribution



# Students/interns: GraphGlove (EMNLP 2020)

## Embedding Words in Non-Vector Space with Unsupervised Graph Learning


Max Ryabinin, Sergei Popov, Liudmila Prokhorenkova, Elena Voita



# Thank you!

Lena Voita

PhD student, Uni Edinburgh & Uni Amsterdam

 lena-voita@hotmail.com

 <https://lena-voita.github.io>

 @lena\_voita

 lena-voita

