DEPARTMENT OF
COMPUTER SCIENCE

Beyond noise: Mitigating the Impact of Fine-grained Semantic Divergences on Neural Machine Translation

Eleftheria Briakou & Marine Carpuat

Supervised Machine Translation (MT)

Typically trained on parallel texts:
Sentences considered as translations of each other

Supervised Machine Translation (MT)

Typically trained on parallel texts:
Sentences considered as translations of each other

votre père est français

Supervised Machine Translation (MT)

Typically trained on parallel texts:
Sentences considered as translations of each other

votre père est français

your father is french

Parallel texts are not always exact translations

votre père est français	your father is french
votre père est français	your parent is french

Parallel texts contain fine-grained semantic divergences

Mostly equivalent parallel texts that contain a small number of divergent tokens

votre père est français	your father is french
votre père est français	your parent is french

Parallel texts are not always exact translations

votre père est français	your father is french
votre père est français	your parent is french
votre père est français	who is your father

Parallel texts contain coarse-grained semantic divergences

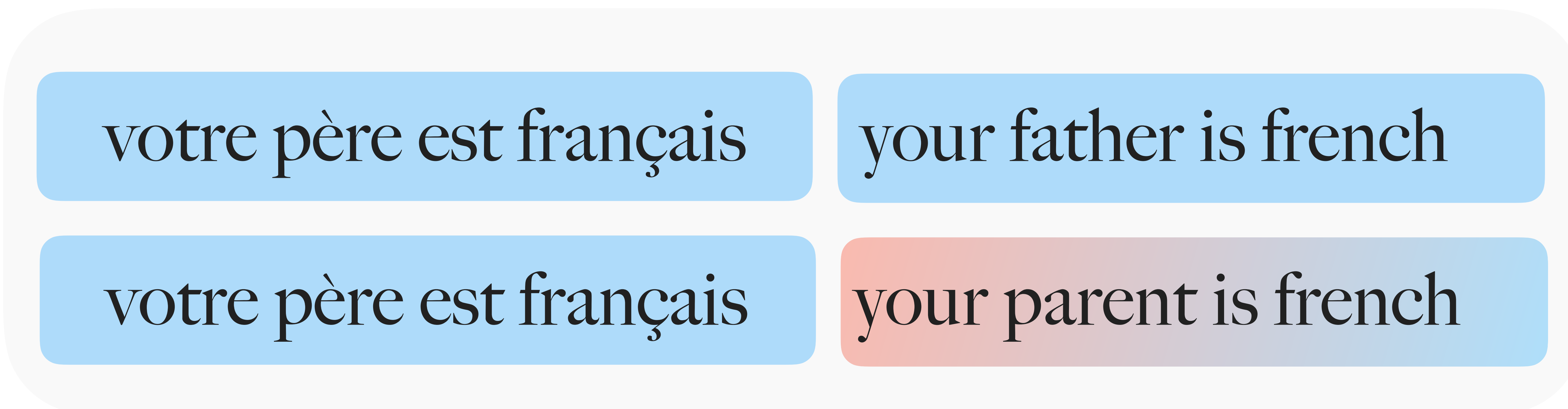
unrelated sentence pairs — noisy training signal

votre père est français	your father is french
votre père est français	your parent is french
votre père est français	who is your father

Coarse-grained semantic divergences are typically excluded from training

votre père est français	your father is french
votre père est français	your parent is french
~~votre père est français~~	~~who is your father~~

Fine-grained semantic divergences are treated as equivalent at MT training



Our work

How do fine-grained divergences impact NMT?

Our work

How do fine-grained divergences impact NMT?

- hurt translation quality
- more repetitive loops
- increase prediction uncertainty

Our work

How do fine-grained divergences impact NMT?

- hurt translation quality
- more repetitive loops
- increase prediction uncertainty

Our work

How do fine-grained divergences impact NMT?

- hurt translation quality
- more repetitive loops
- increase prediction uncertainty

Our work

How do fine-grained divergences impact NMT?

- hurt translation quality
- more repetitive loops
- increase prediction uncertainty

How can we mitigate their negative impact?

Our work

How do fine-grained divergences impact NMT?

- hurt translation quality
- more repetitive loops
- increase prediction uncertainty

How can we mitigate their negative impact?

- by encoding divergences as token factors

Divergences matter for NMT because they yield unreliable training signals

$$J(\theta) = \sum_{n=1}^{N} \sum_{t=1}^{T} \log p(y_t^{(n)} \mid y_{<t}^{(n)}, x^{(n)}; \theta)$$

votre père est français

your parent is french

Divergences matter for NMT because they yield unreliable training signals

$$J(\theta) = \sum_{n=1}^{N} \sum_{t=1}^{T} \log p(y_t^{(n)} \mid y_{<t}^{(n)}, \boxed{x^{(n)};}\, \theta)$$

votre père est français

your parent is french

Divergences matter for NMT because they yield unreliable training signals

$$J(\theta) = \sum_{n=1}^{N} \sum_{t=1}^{T} \log p(y_t^{(n)} \mid y_{<t}^{(n)}, x^{(n)}; \theta)$$

$$t = 1$$

votre père est français

your parent is french

Divergences matter for NMT because they yield unreliable training signals

$$J(\theta) = \sum_{n=1}^{N} \sum_{t=1}^{T} \log p(y_t^{(n)} \mid y_{<t}^{(n)}, x^{(n)}; \theta)$$

$$t = 2$$

votre père est français

your parent is french

Divergences matter for NMT because they yield unreliable training signals

$$J(\theta) = \sum_{n=1}^{N} \sum_{t=1}^{T} \log p(\boxed{y_t^{(n)}} \mid \boxed{y_{<t}^{(n)}}, x^{(n)}; \theta)$$

$$t = 3$$

votre père est français

your parent is french

Divergences matter for NMT because they yield unreliable training signals

$$J(\theta) = \sum_{n=1}^{N} \sum_{t=1}^{T} \log p(y_t^{(n)} \mid y_{<t}^{(n)}, x^{(n)}; \theta)$$

$$t = 4$$

votre père est français

your parent is french

How do fine-grained divergences impact NMT?

Controlled analysis on artificial divergences

How do fine-grained divergences impact NMT?

Controlled analysis on artificial divergences

Experimental Setting

How do fine-grained divergences impact NMT?

Controlled analysis on artificial divergences

Experimental Setting

- Training bitext : WikiMatrix (mined)
- Test set : TED
- Language-pair : French → English
- NMT architecture : Transformer

How do fine-grained divergences impact NMT?

Controlled analysis on artificial divergences

Experimental Setting

- Training bitext : WikiMatrix (mined)
- Test set : TED
- Language-pair : French → English
- NMT architecture : Transformer

Measuring the impact of synthetic divergences on NMT

EQUIVALENT

ils vous demandent votre aide	they are asking your help

Measuring the impact of synthetic divergences on NMT

EQUIVALENT	
ils vous demandent votre aide	they are asking your help

PHRASE DELETION	
ils vous demandent votre aide	they are asking

Measuring the impact of synthetic divergences on NMT

EQUIVALENT

ils vous demandent votre aide	they are asking your help

PHRASE DELETION

ils vous demandent votre aide	they are asking

LEXICAL SUBSTITUTION

ils vous demandent votre aide	they are asking your mercy

Measuring the impact of synthetic divergences on NMT

EQUIVALENT

ils vous demandent votre aide	they are asking your help

PHRASE DELETION

ils vous demandent votre aide	they are asking

LEXICAL SUBSTITUTION

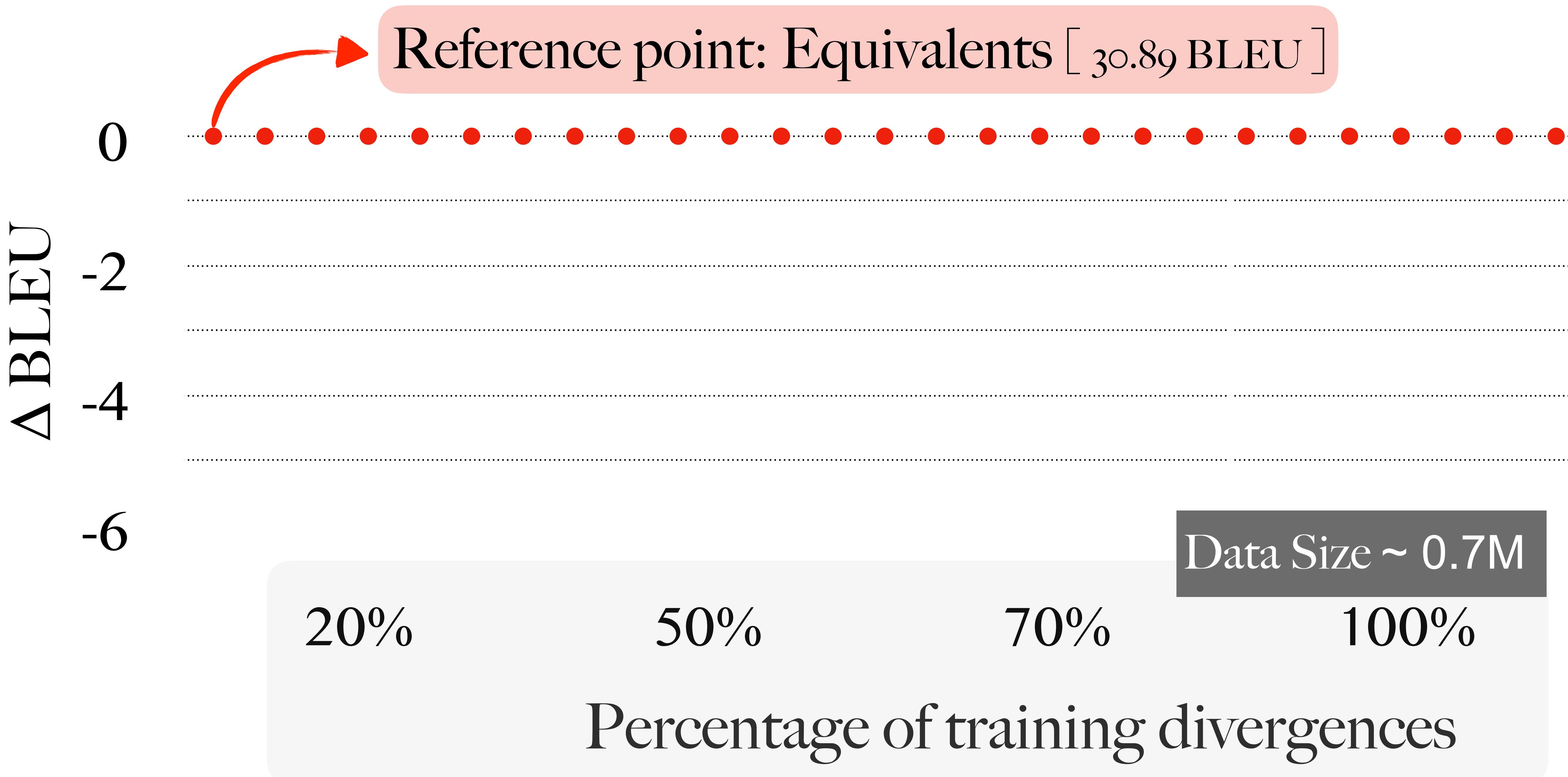
ils vous demandent votre aide	they are asking your mercy

PHRASE REPLACEMENT

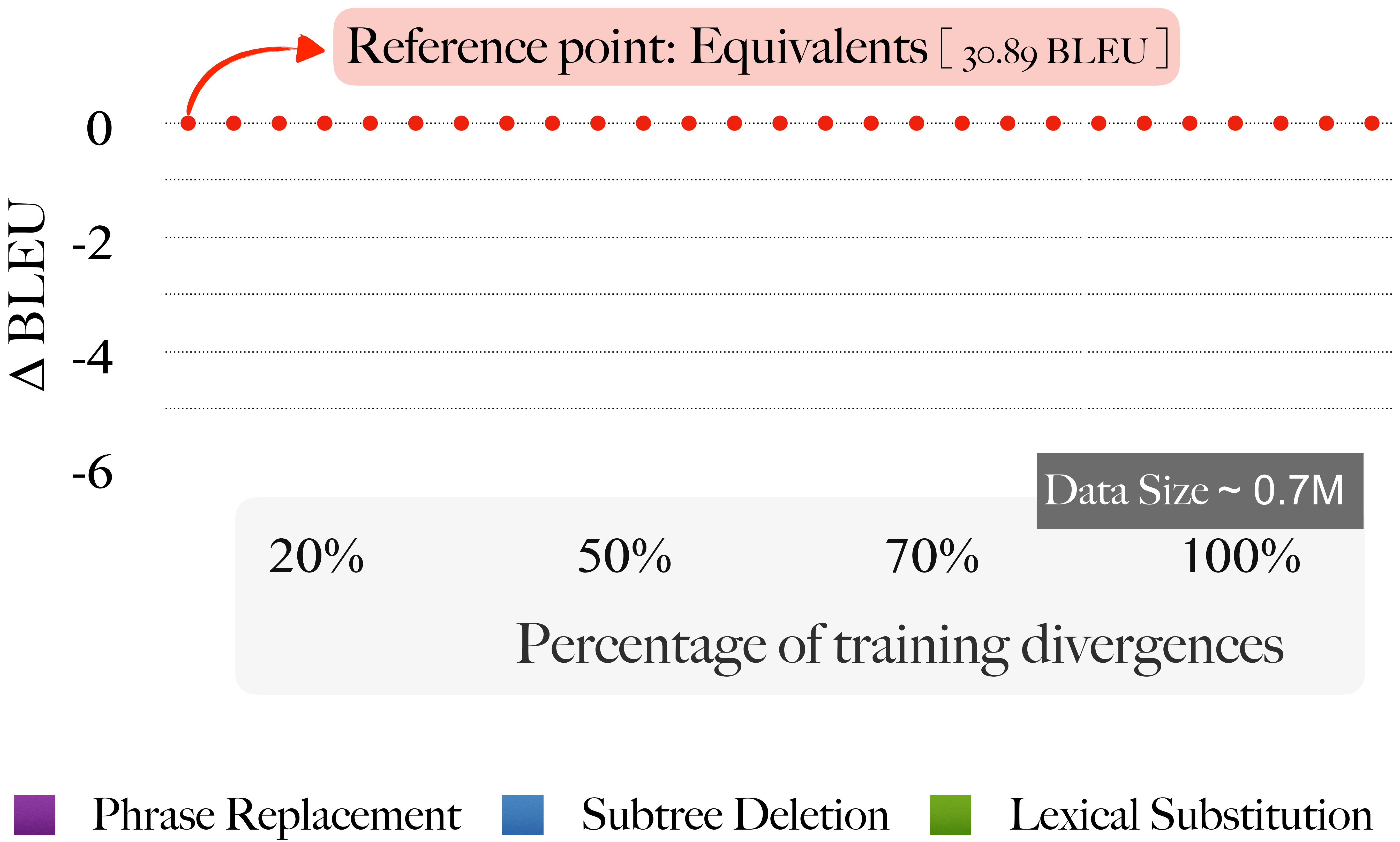
ils vous demandent votre aide	they were ignoring his help

Fine-grained Divergences: Impact on BLEU

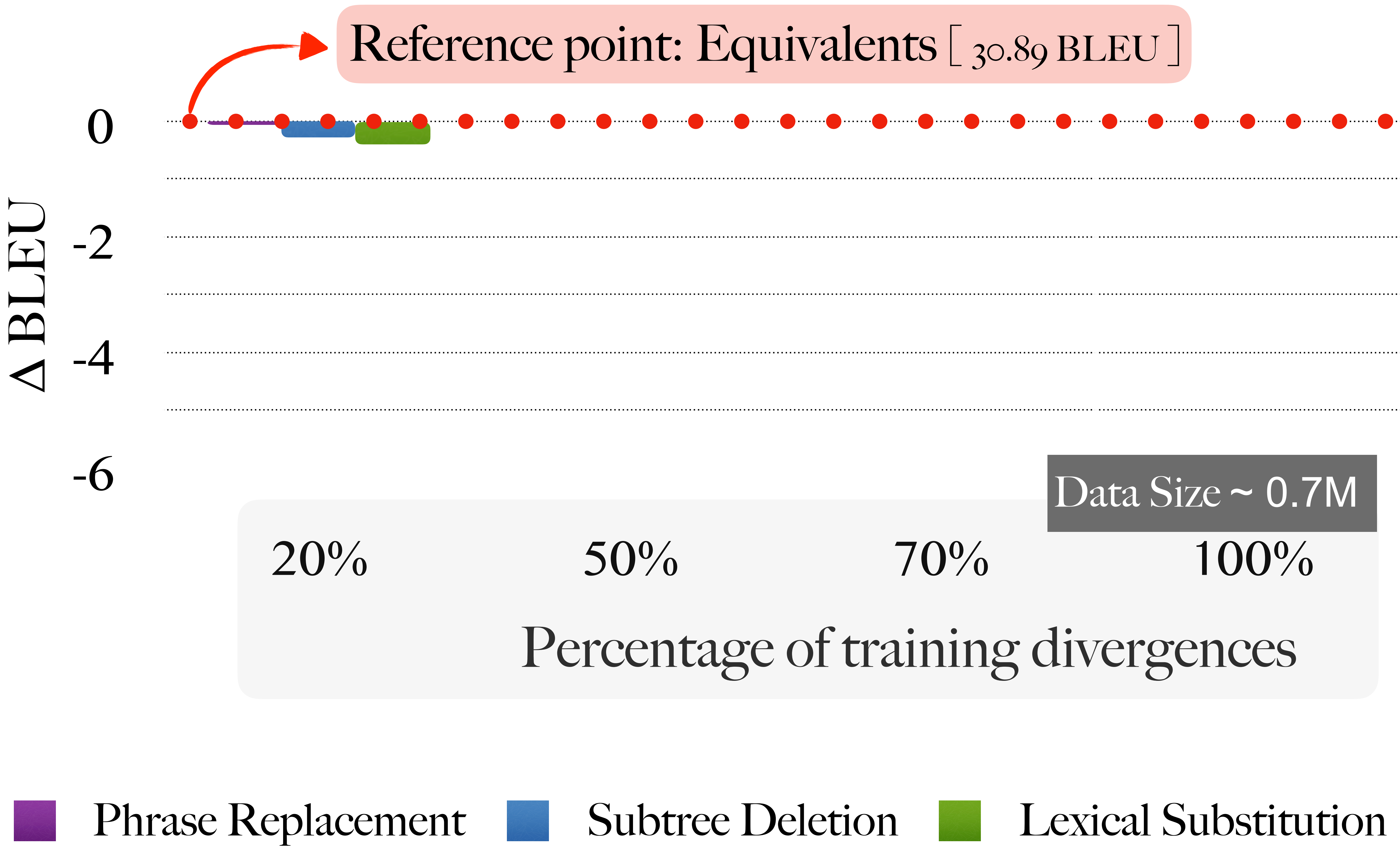
Fine-grained Divergences: Impact on BLEU



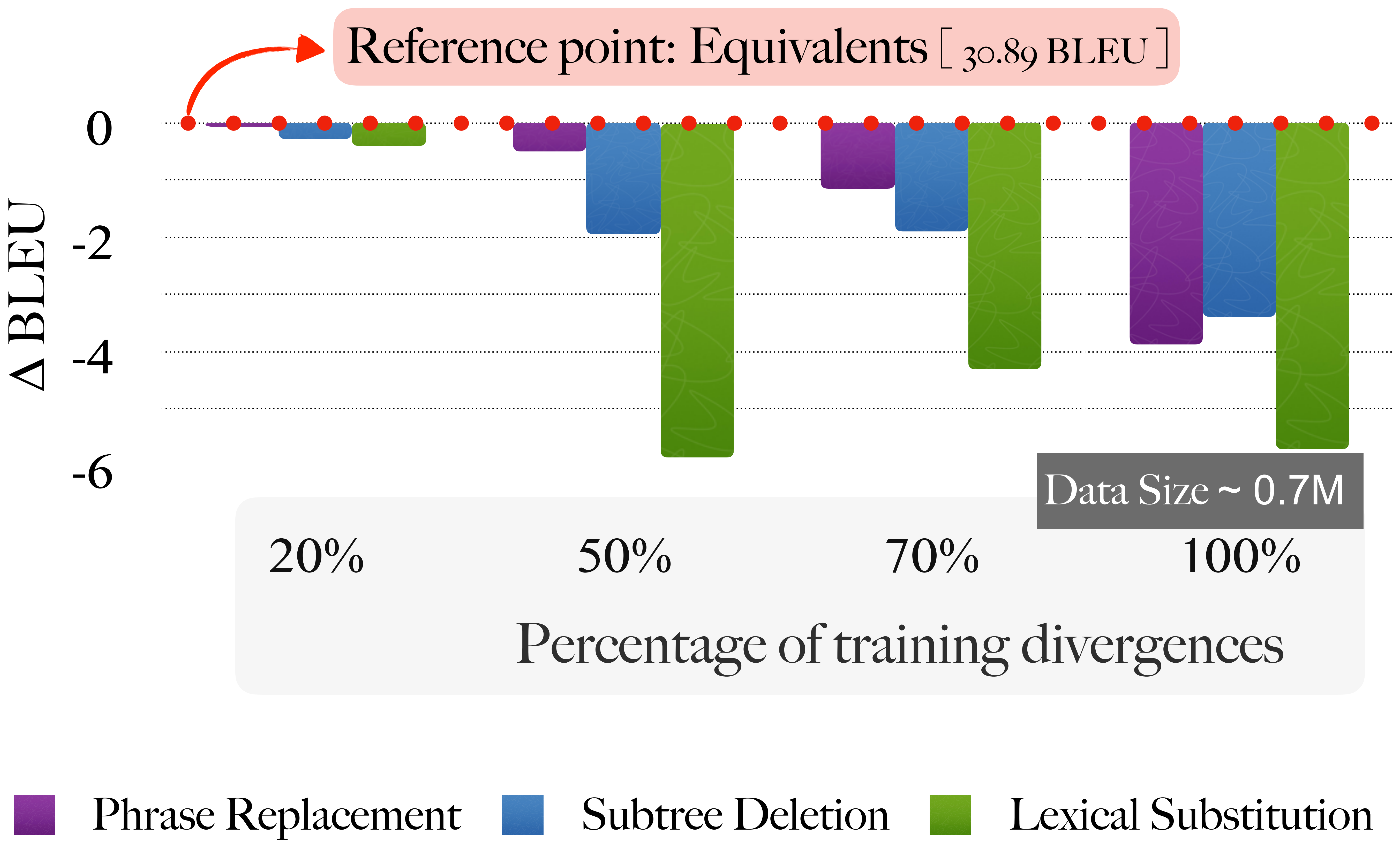
Fine-grained Divergences: Impact on BLEU

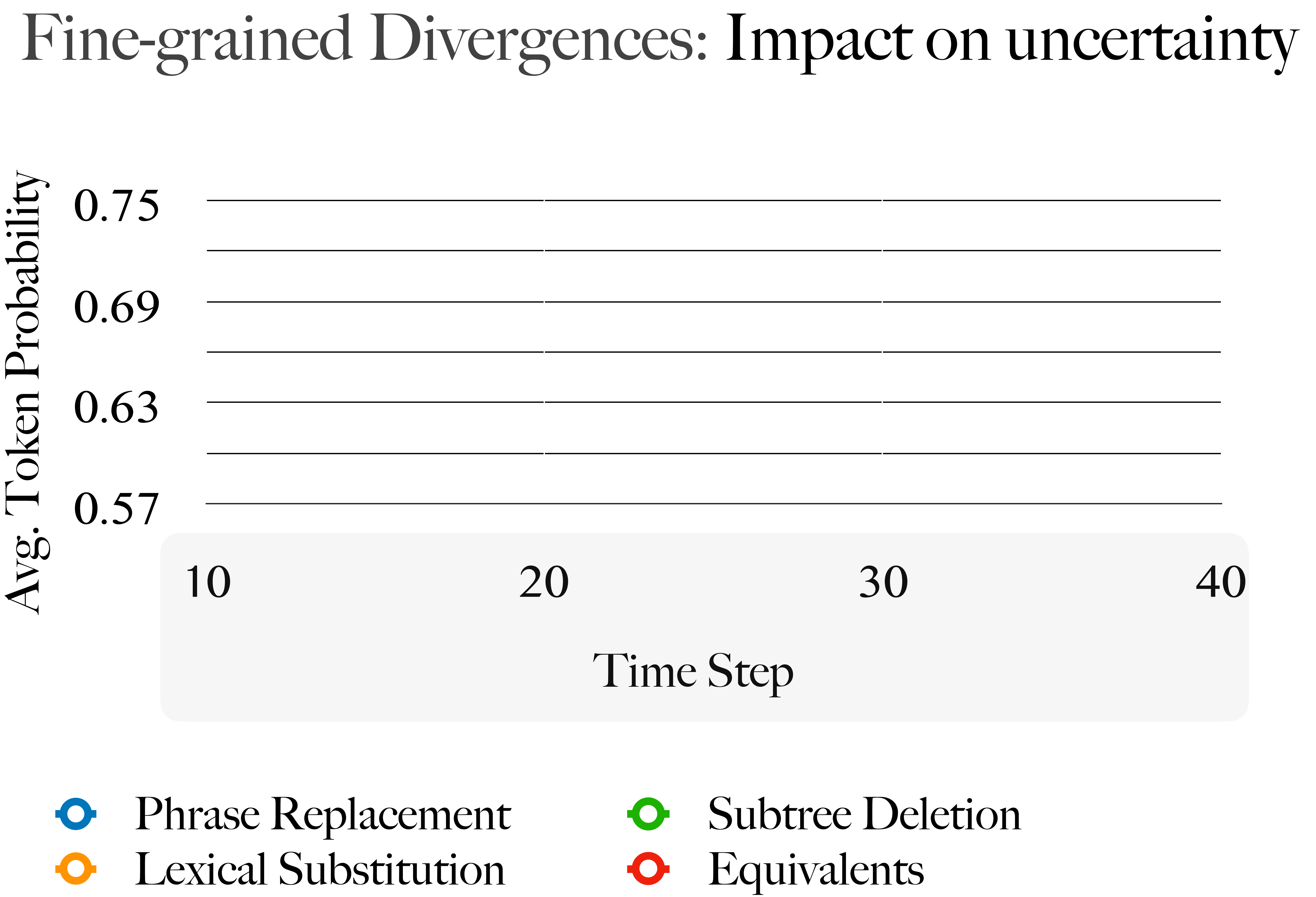


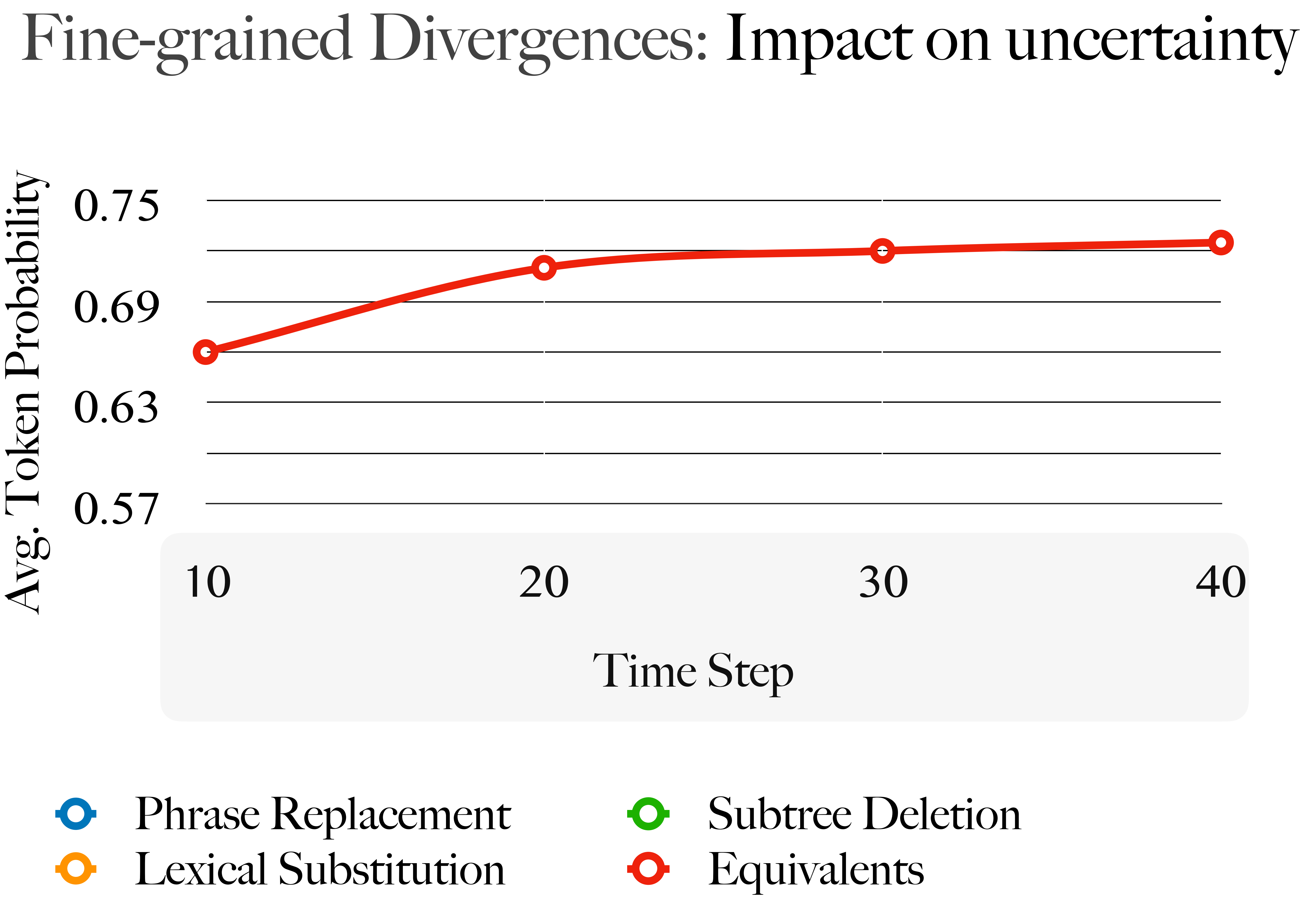
Fine-grained Divergences have small impact on BLEU when equivalents overwhelm training data

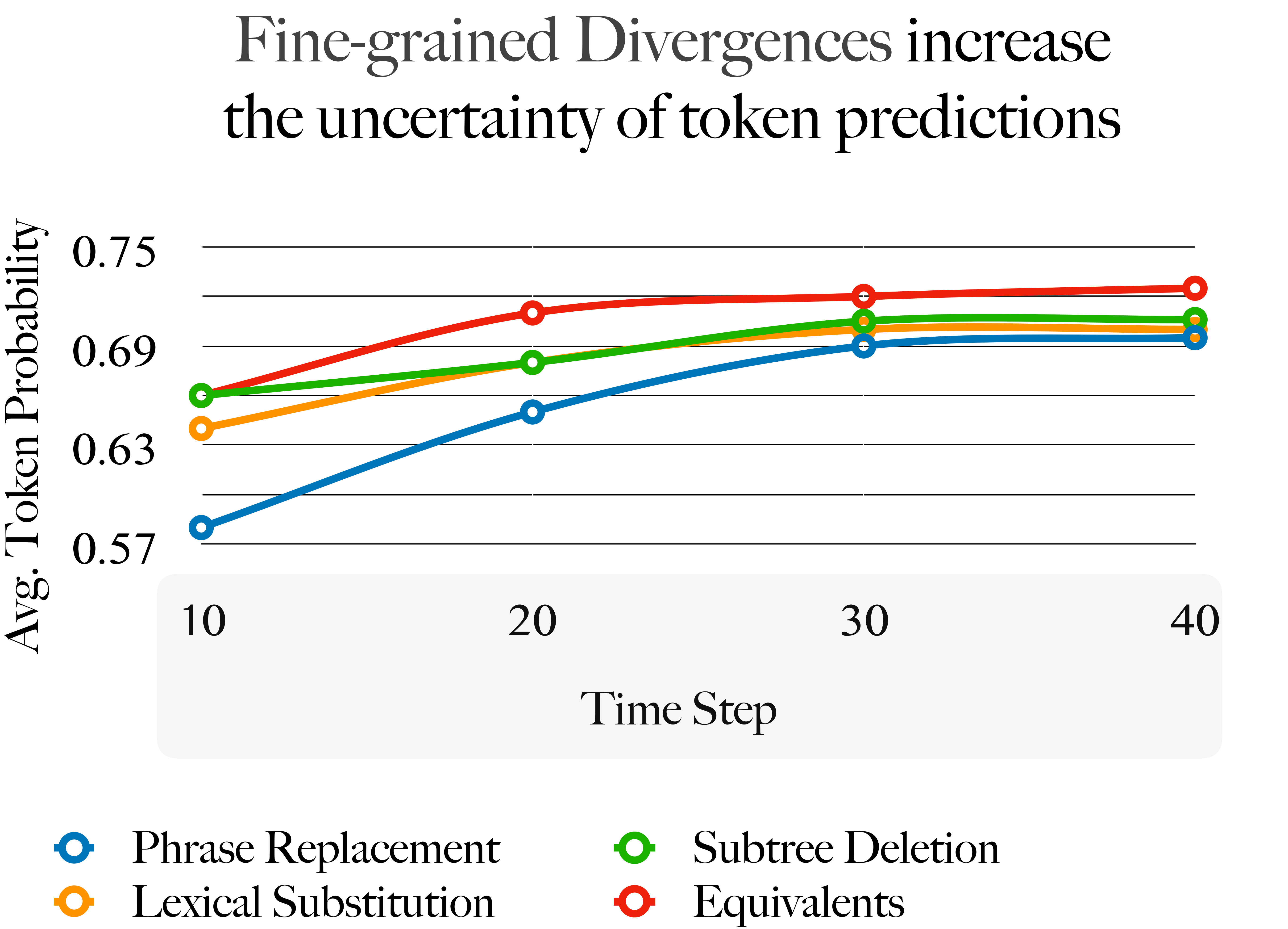


Fine-grained Divergences degrade BLEU when they overwhelm the training data



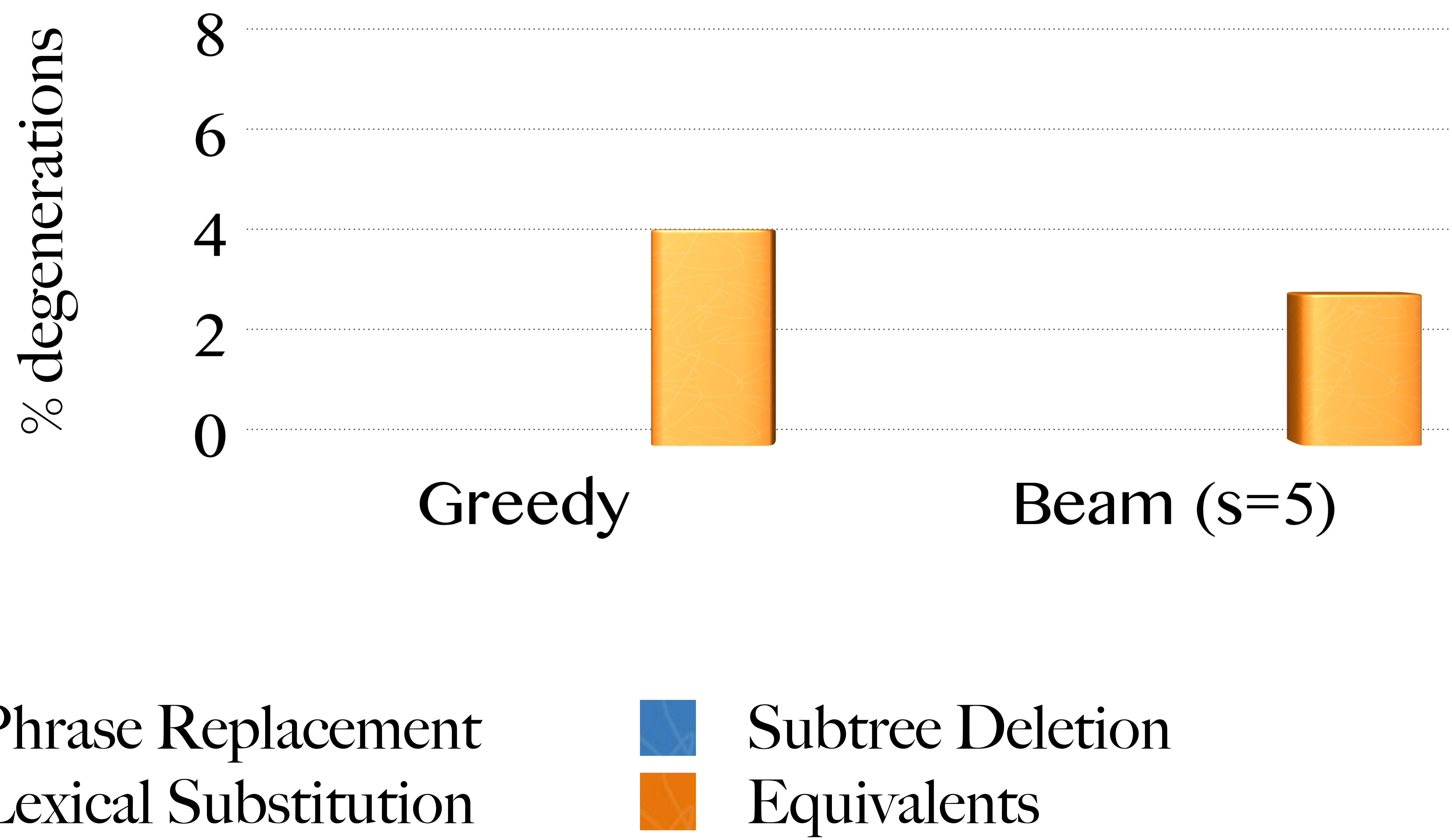

Fine-grained Divergences: Impact on uncertainty
Avg. Token Probability
0.75
0.69
0.63
0.57
10
20
30
40
Time Step
Phrase Replacement
Lexical Substitution
Subtree Deletion
Equivalents


Fine-grained Divergences: Impact on uncertainty
Avg. Token Probability
0.75
0.69
0.63
0.57
10
20
30
40
Time Step
Phrase Replacement
Subtree Deletion
Lexical Substitution
Equivalents


Fine-grained Divergences increase
the uncertainty of token predictions
Avg. Token Probability
0.75
0.69
0.63
0.57
10
20
30
40
Time Step
Phrase Replacement
Subtree Deletion
Lexical Substitution
Equivalents

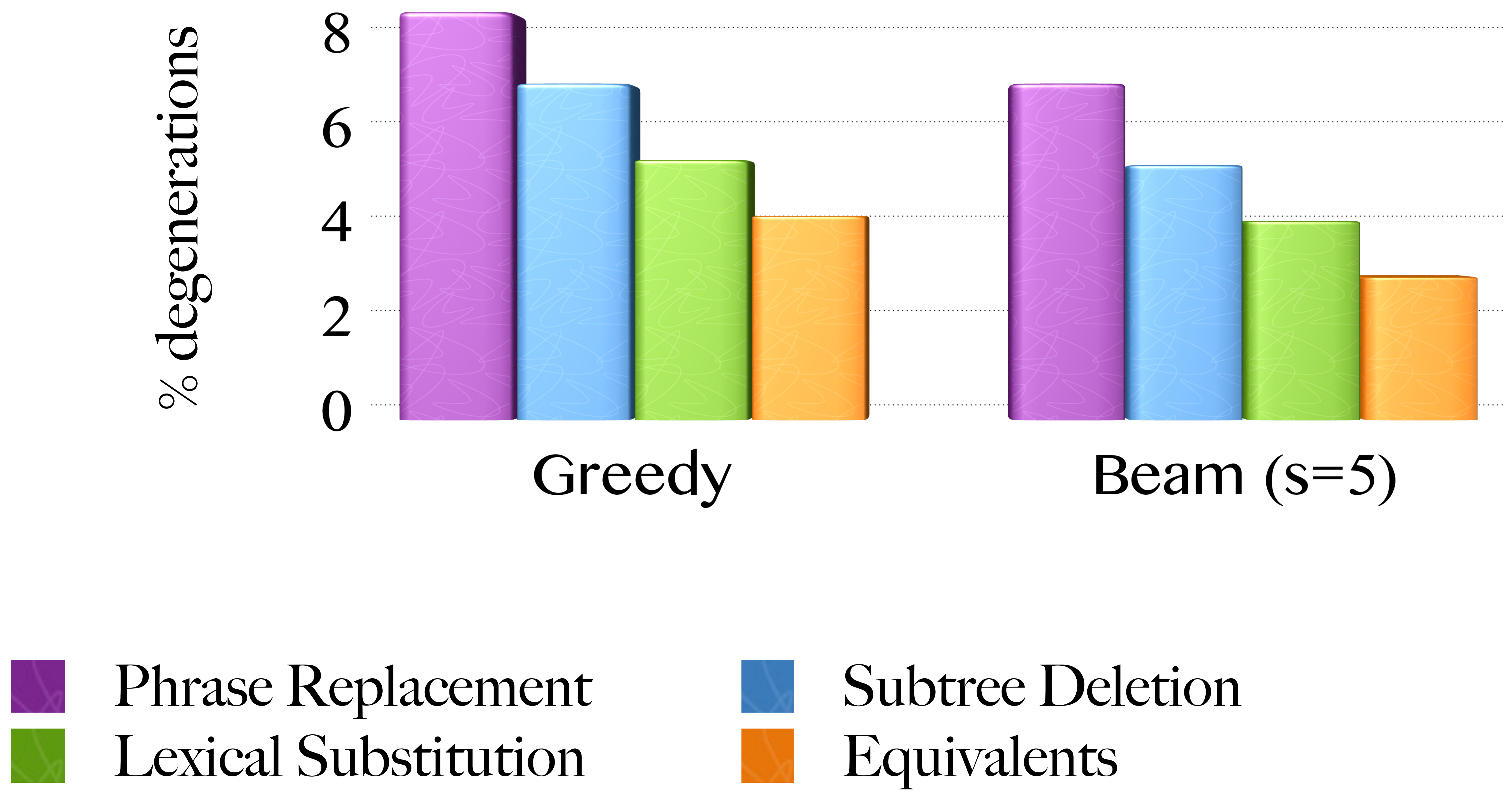
Fine-grained Divergences: Impact on degenerated hypotheses

i.e., *"I've never studied sculpture, engineering and architecture, and the engineering and architecture"*



Fine-grained Divergences increase the frequency of degenerated hypotheses

i.e., *"I've never studied sculpture, engineering and architecture, and the engineering and architecture"*



Our work

How do fine-grained divergences impact NMT?

- hurt translation quality
- more repetitive loops
- increase prediction uncertainty

How can we mitigate their negative impact?

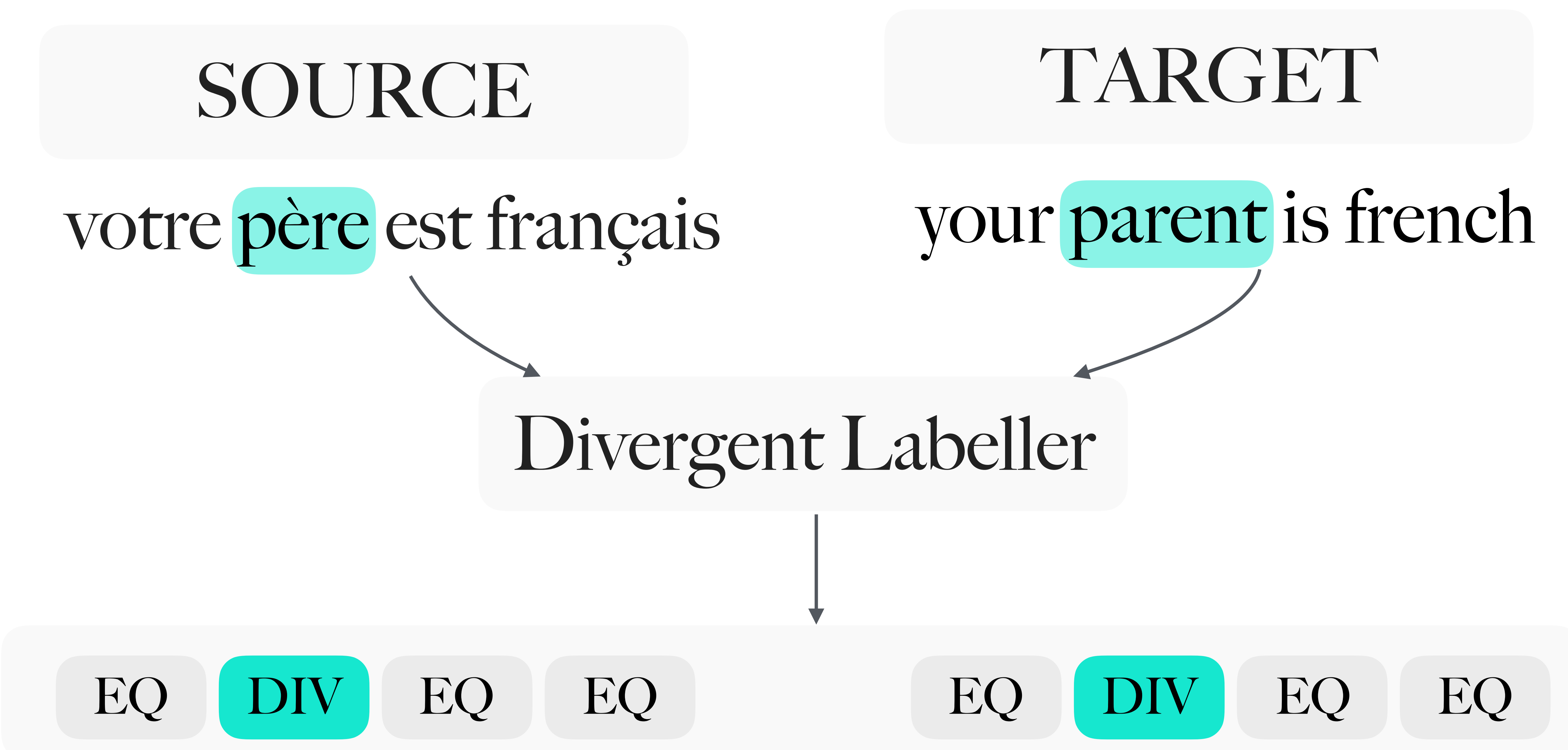
- by encoding divergences as token factors

DIV-FACTORS: Inform NMT training of divergent tokens

SOURCE	TARGET
votre père est français	your parent is french

DIV-FACTORS: Inform NMT training of divergent tokens

DIV-FACTORS: Inform NMT training of divergent tokens

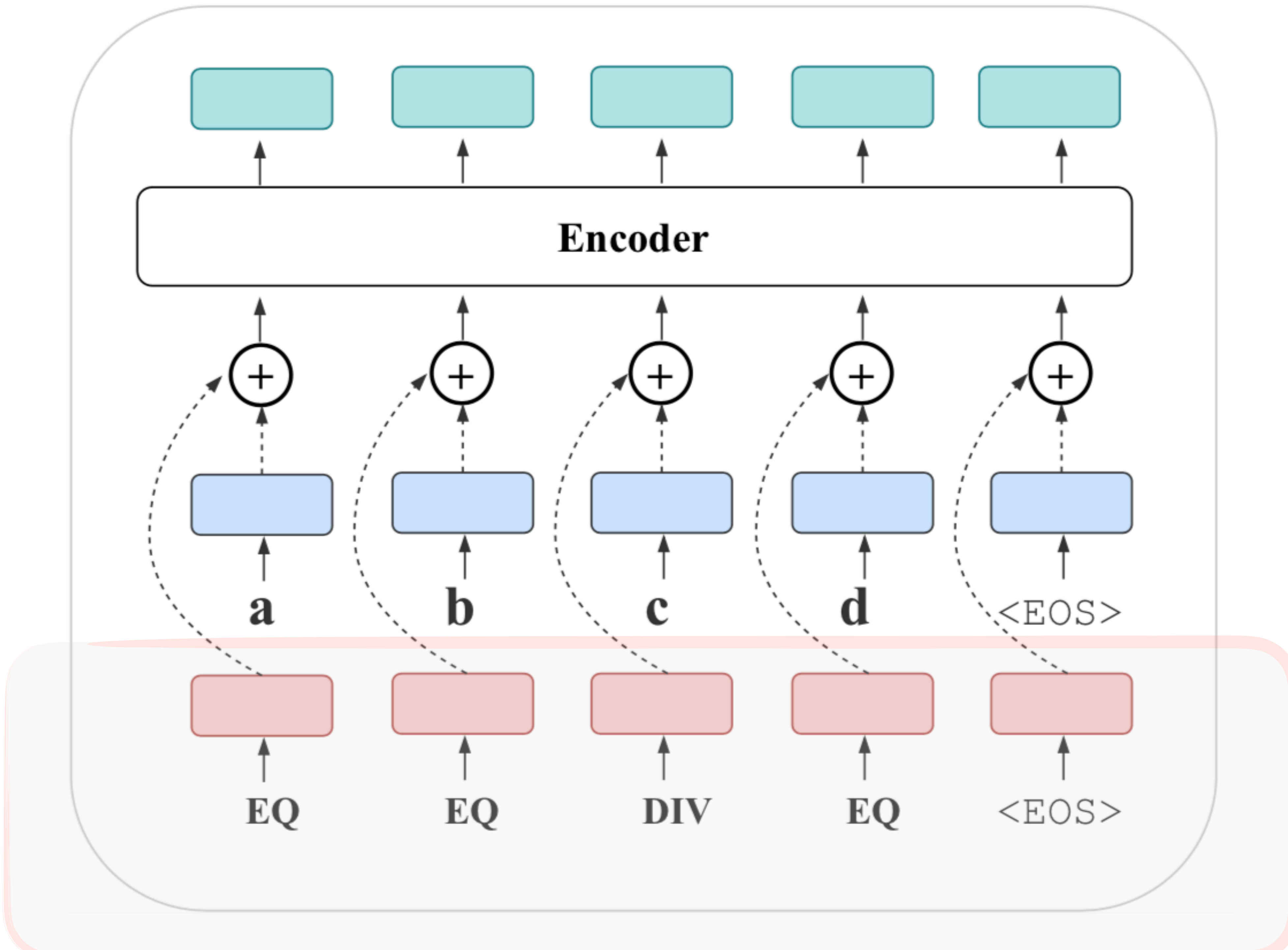


DIV-FACTORS: Inform NMT training of divergent tokens

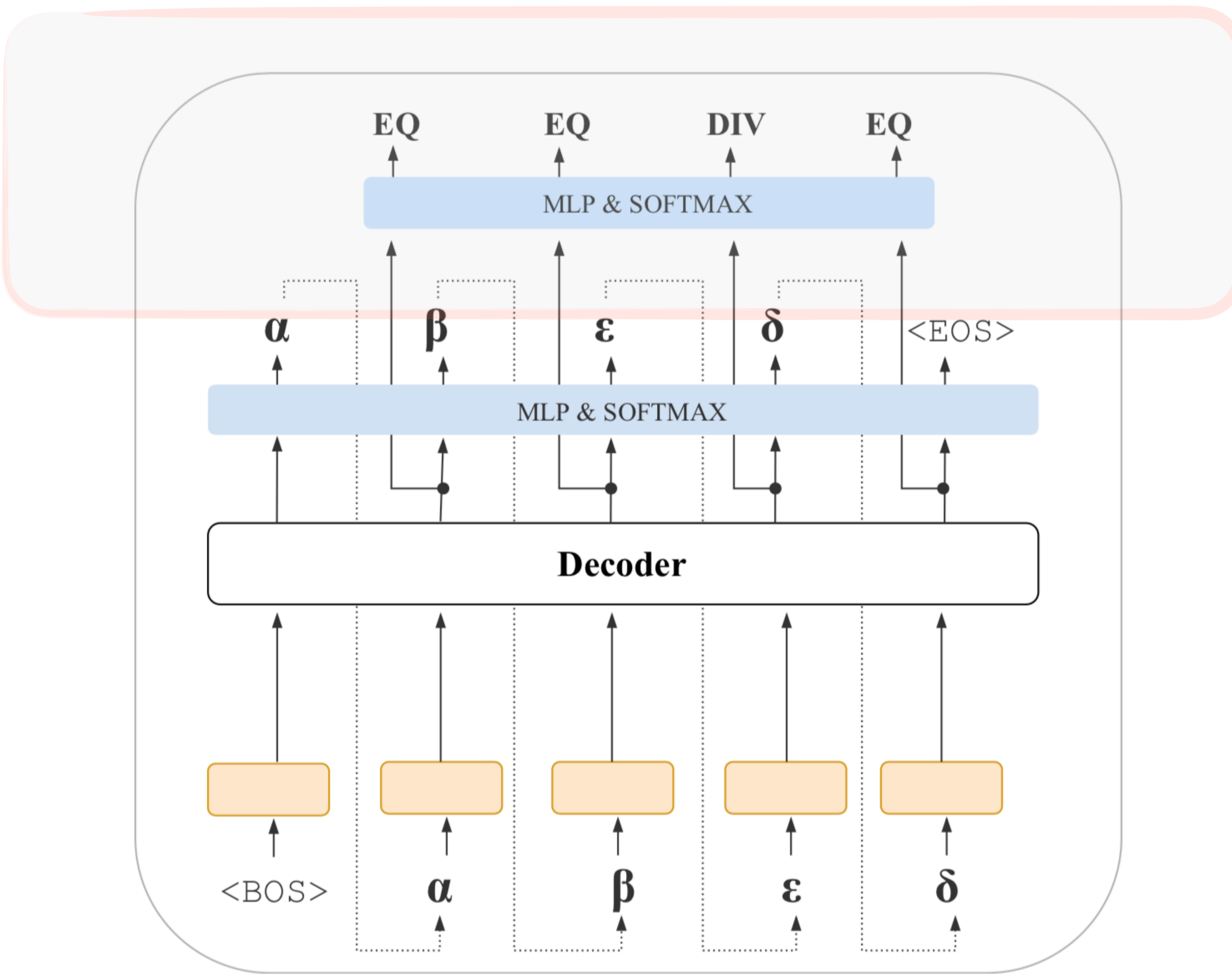
SOURCE			
votre	père	est	français
EQ	DIV	EQ	EQ

TARGET			
your	parent	is	french
EQ	DIV	EQ	EQ

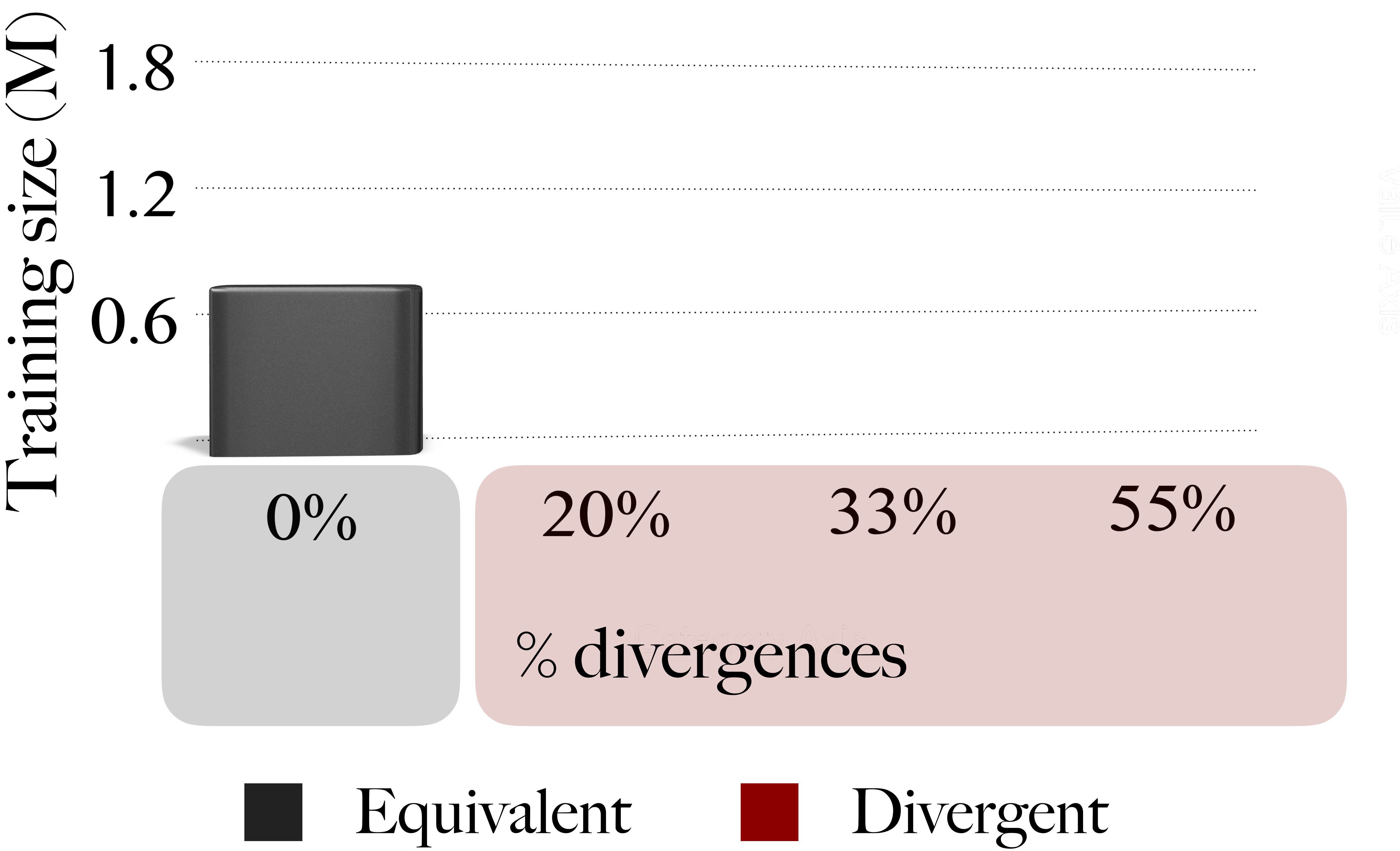
Source-side factors: divergent tags are encoded as additional features



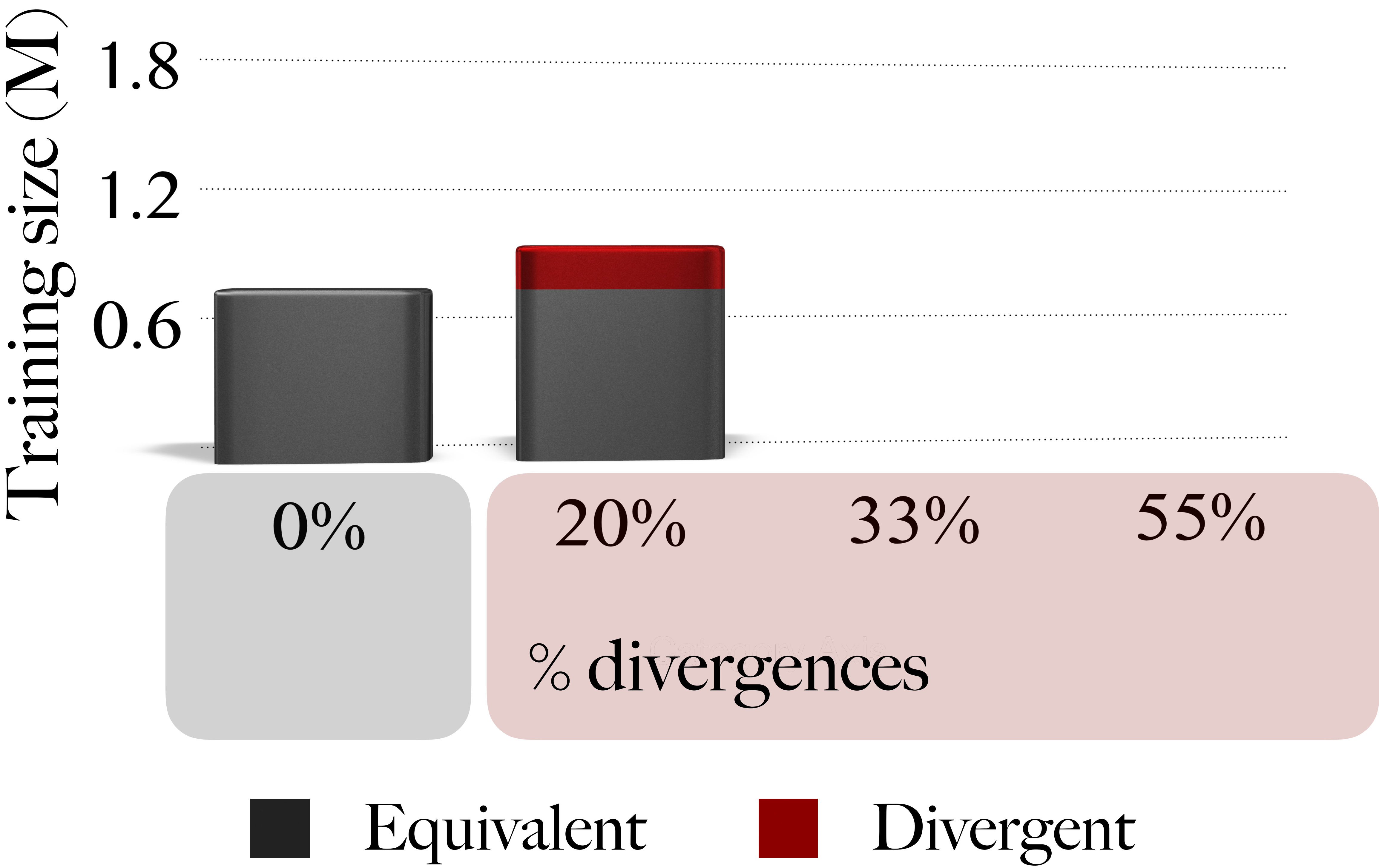
Target-side factors: divergent tags are generated additional sequence



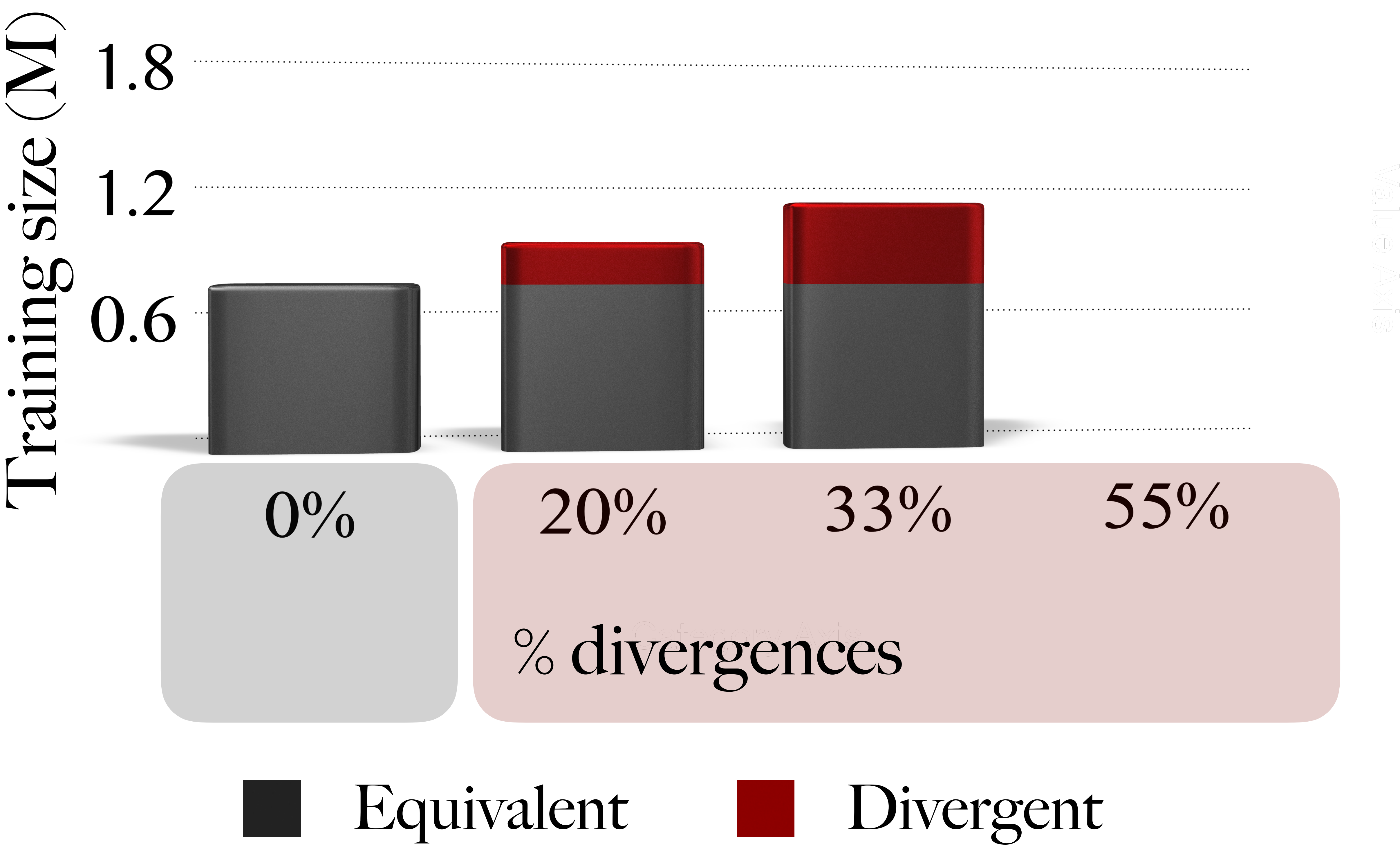
Mitigating the impact of divergences: Experimental Setup



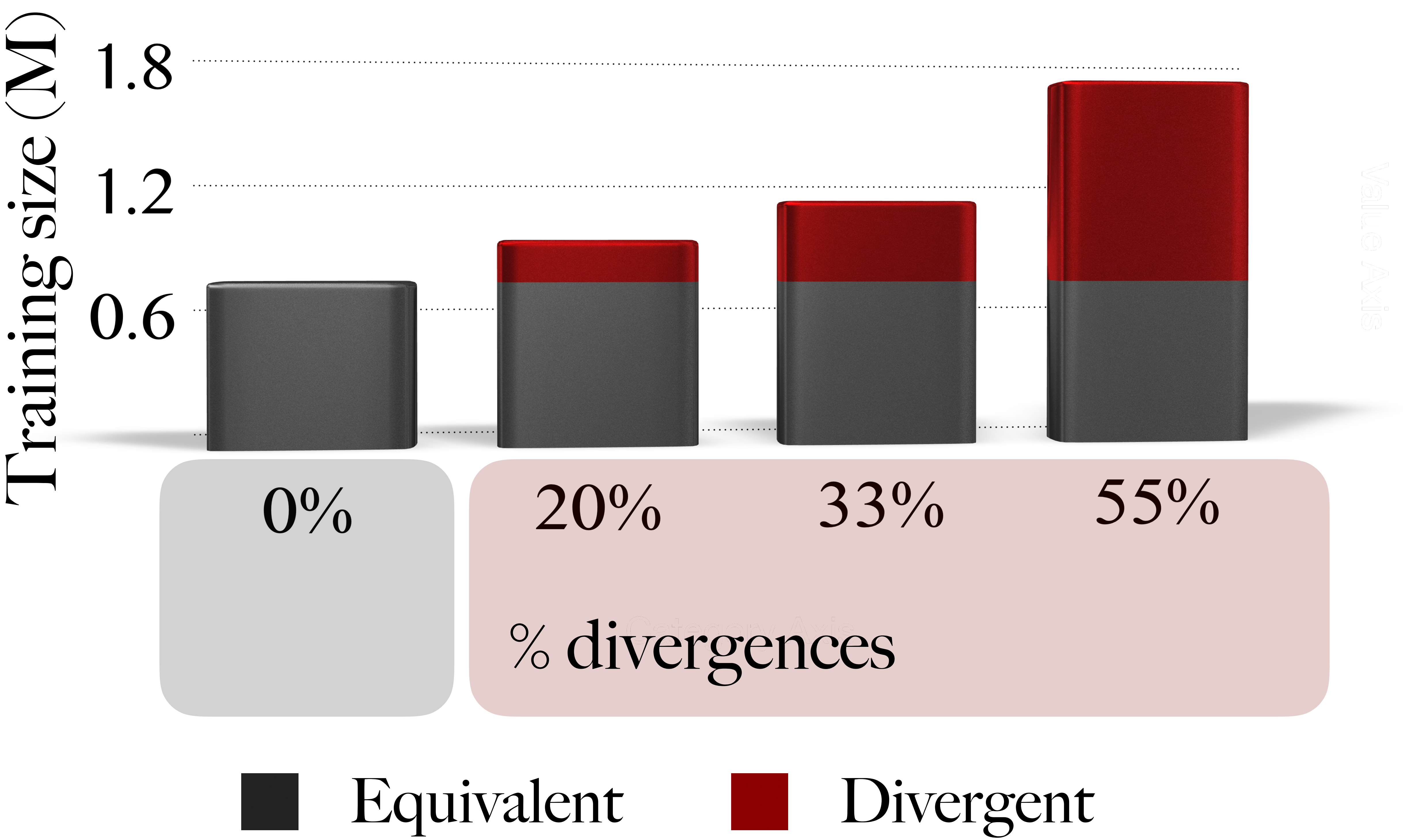
Mitigating the impact of divergences: Experimental Setup



Mitigating the impact of divergences: Experimental Setup



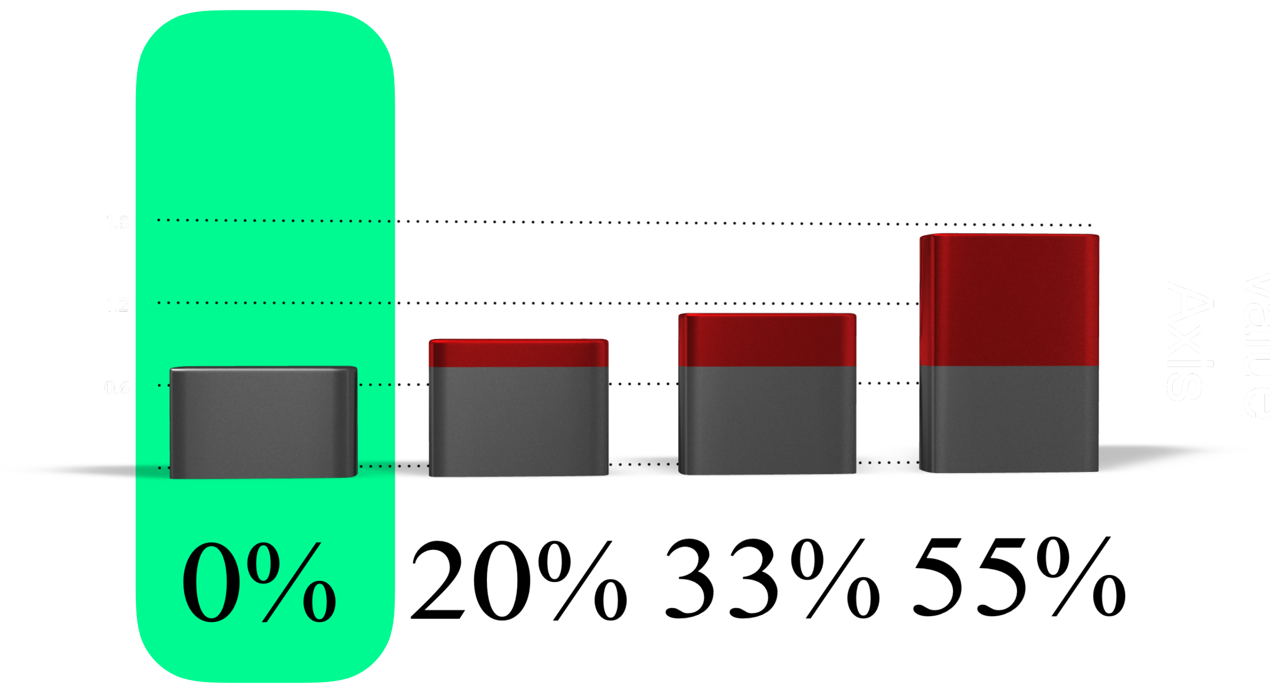
Mitigating the impact of divergences: Experimental Setup



Mitigating the impact of divergences: Experimental Setup

Models

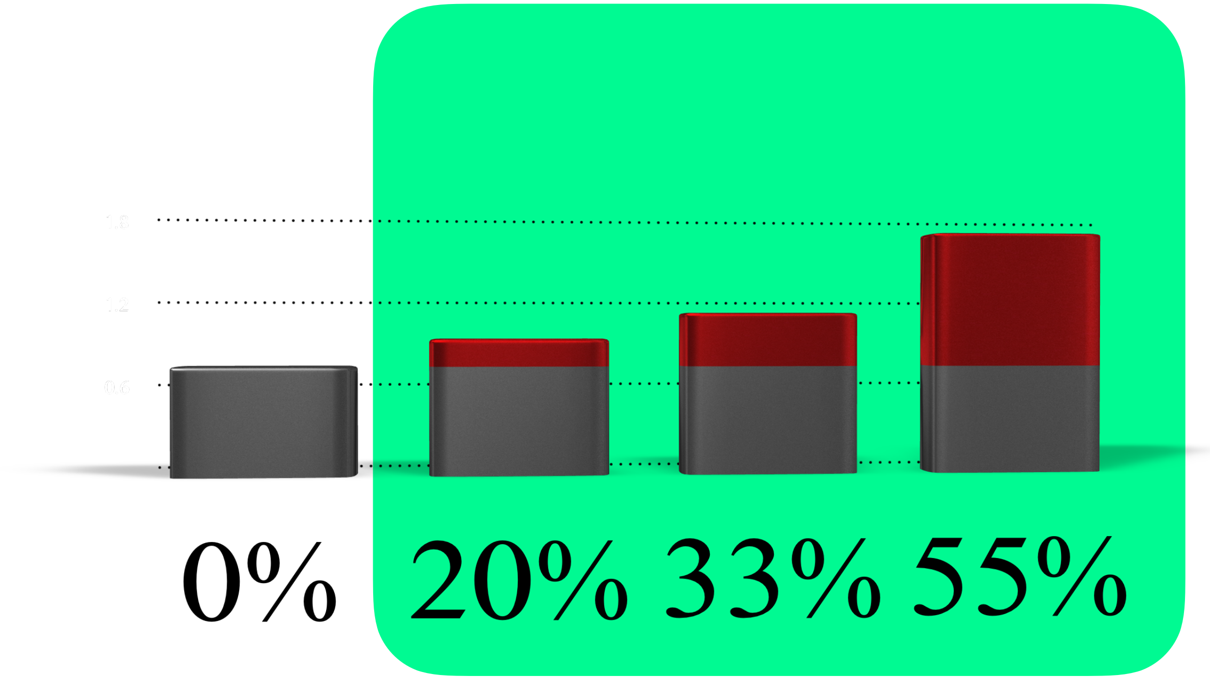
- Equivalents



Mitigating the impact of divergences: Experimental Setup

<u>Models</u>

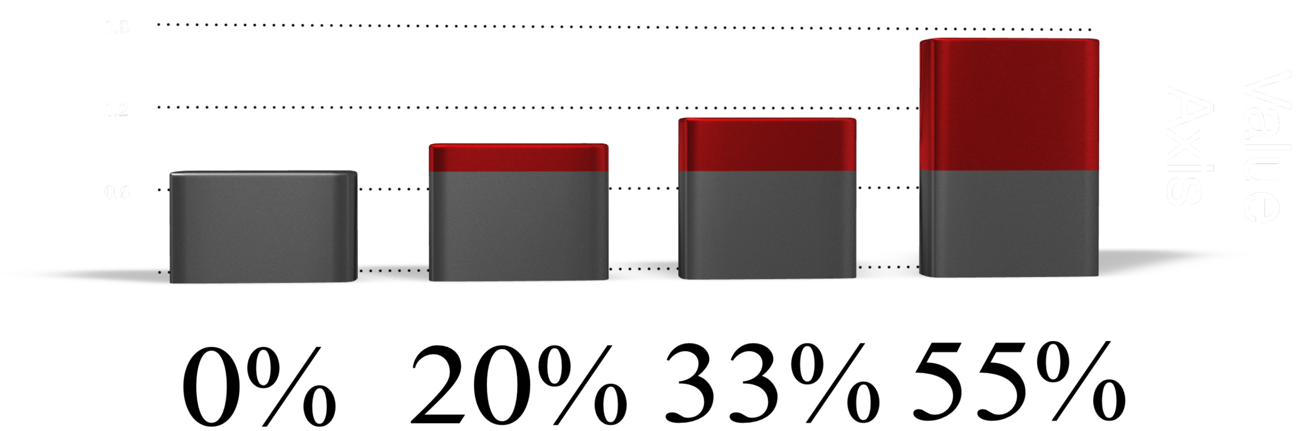
- Equivalents
- DIV-AGNOSTIC
- DIV-FACTORS



Mitigating the impact of divergences: Experimental Setup

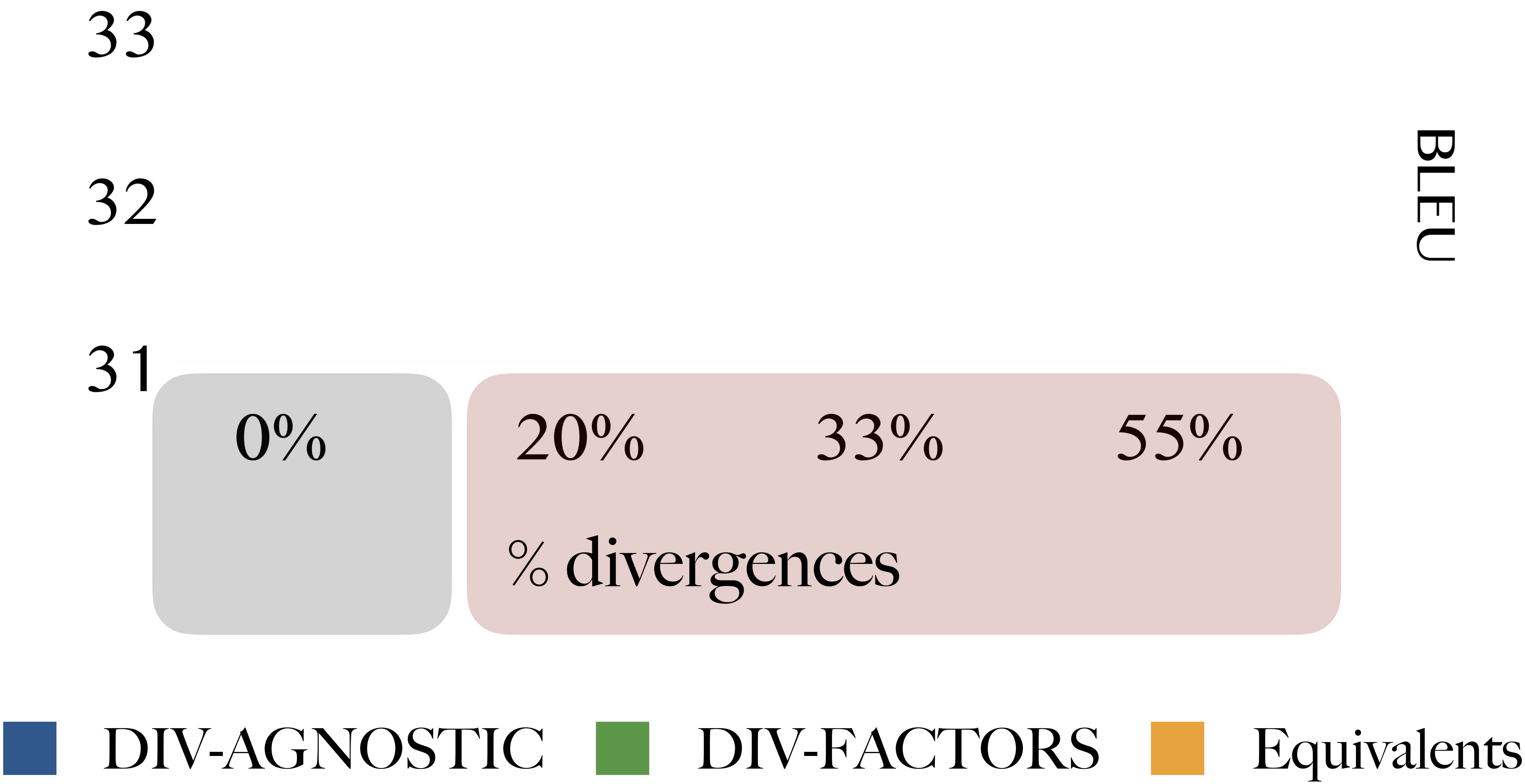
Models

- Equivalents
- DIV-AGNOSTIC
- DIV-FACTORS

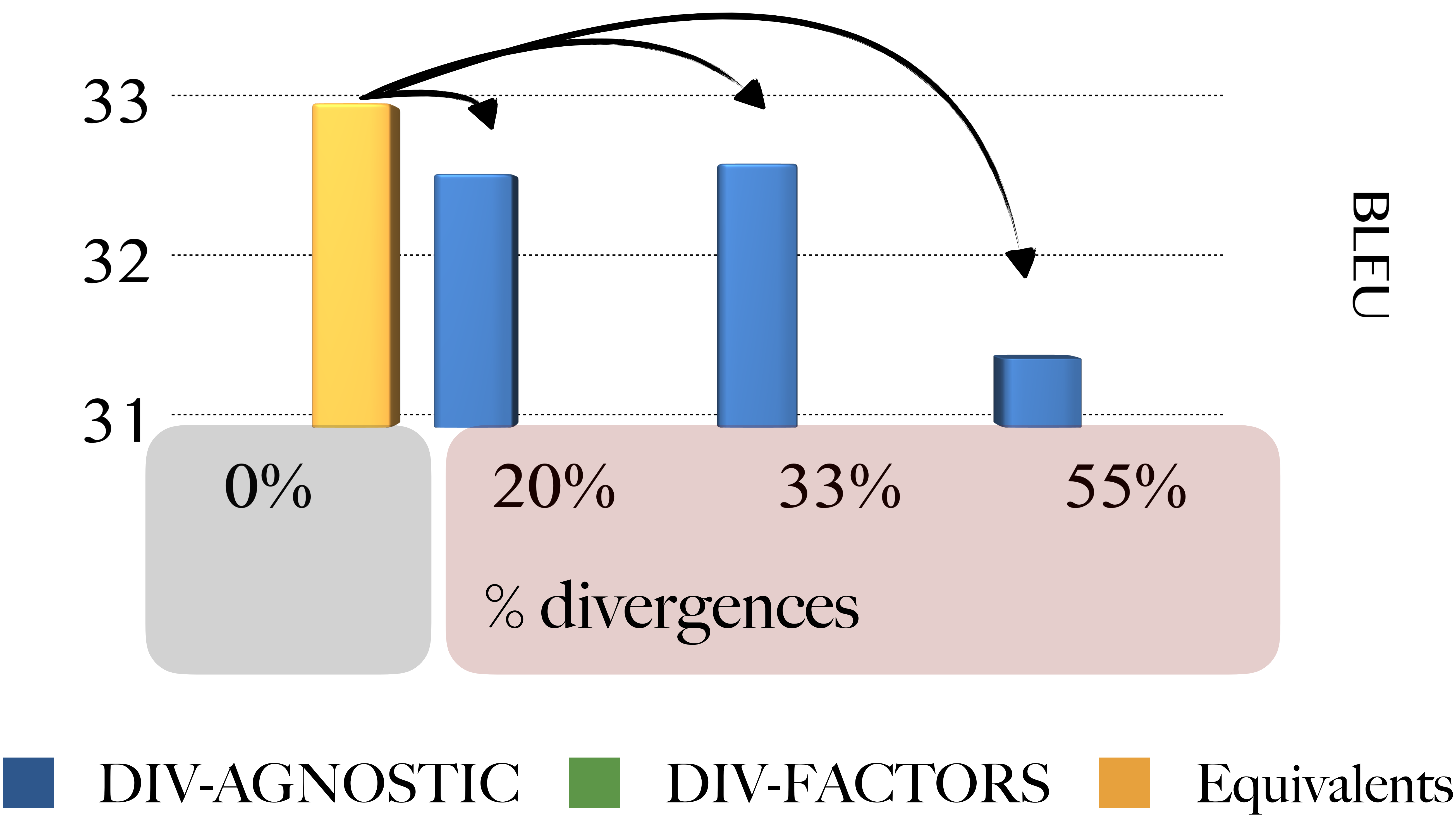


- Training bitext : WikiMatrix (mined)
- Test set : TED
- Language-pair : French ↔ English
- NMT architecture : Transformer

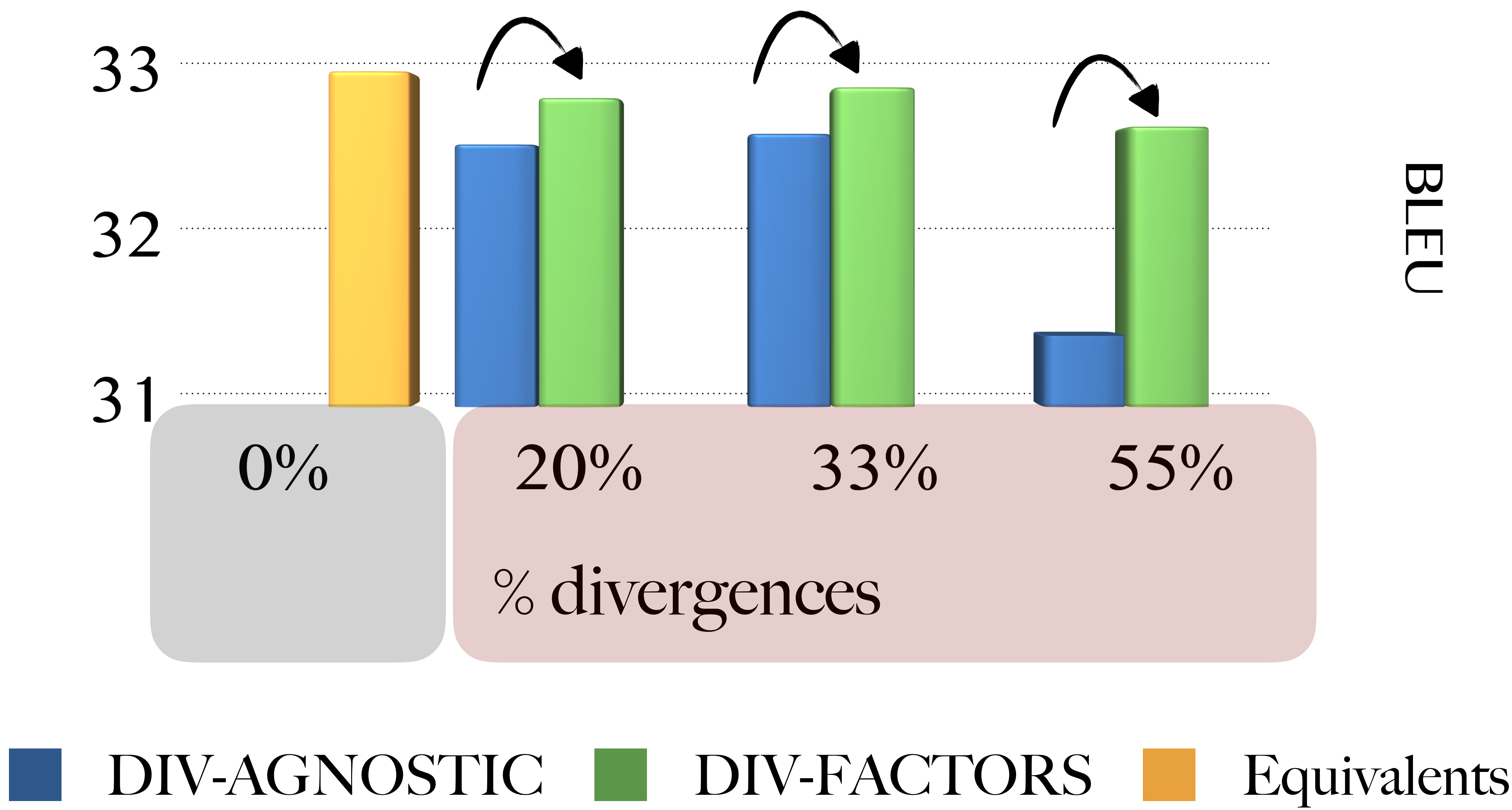
Divergences decrease translation quality

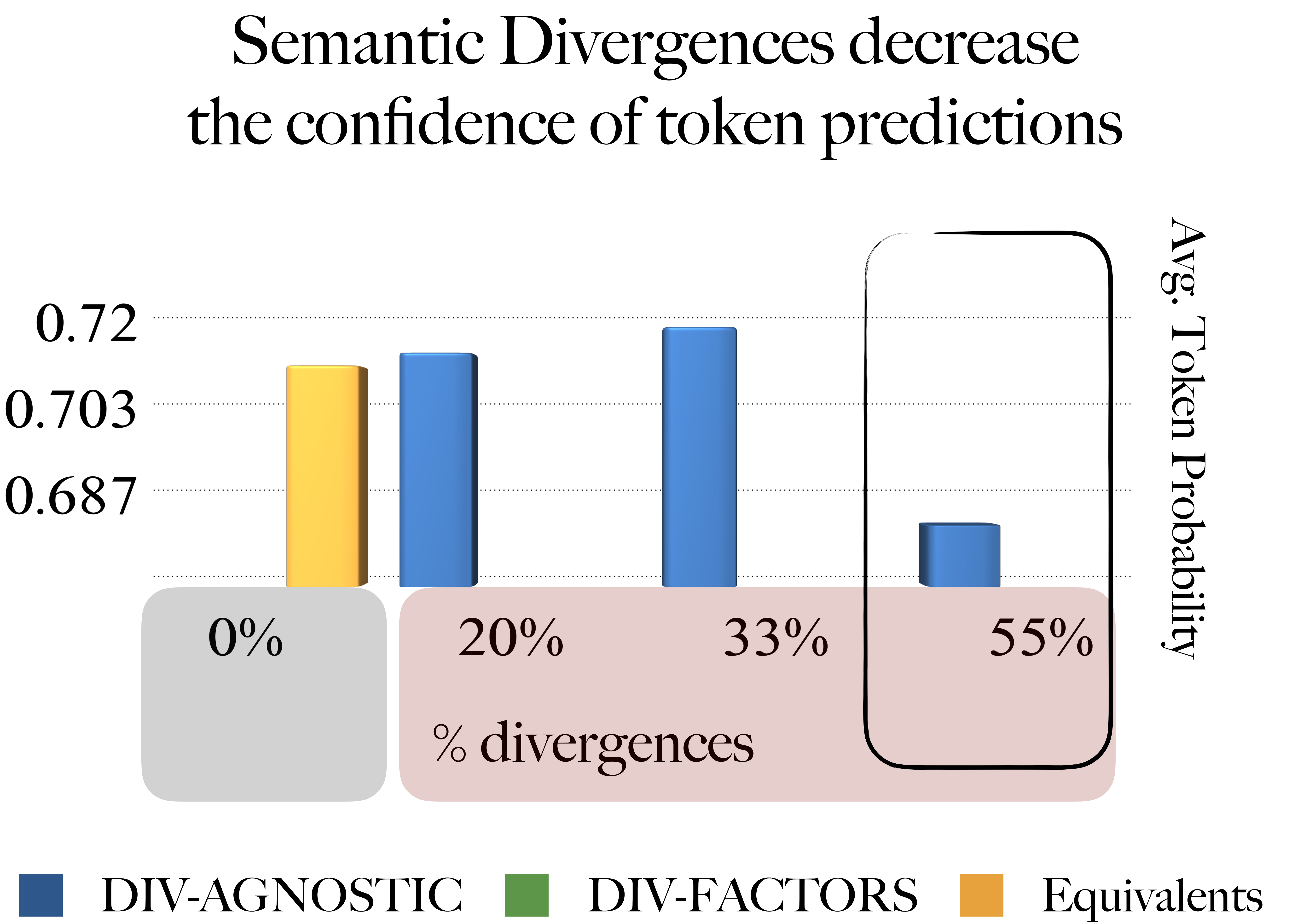


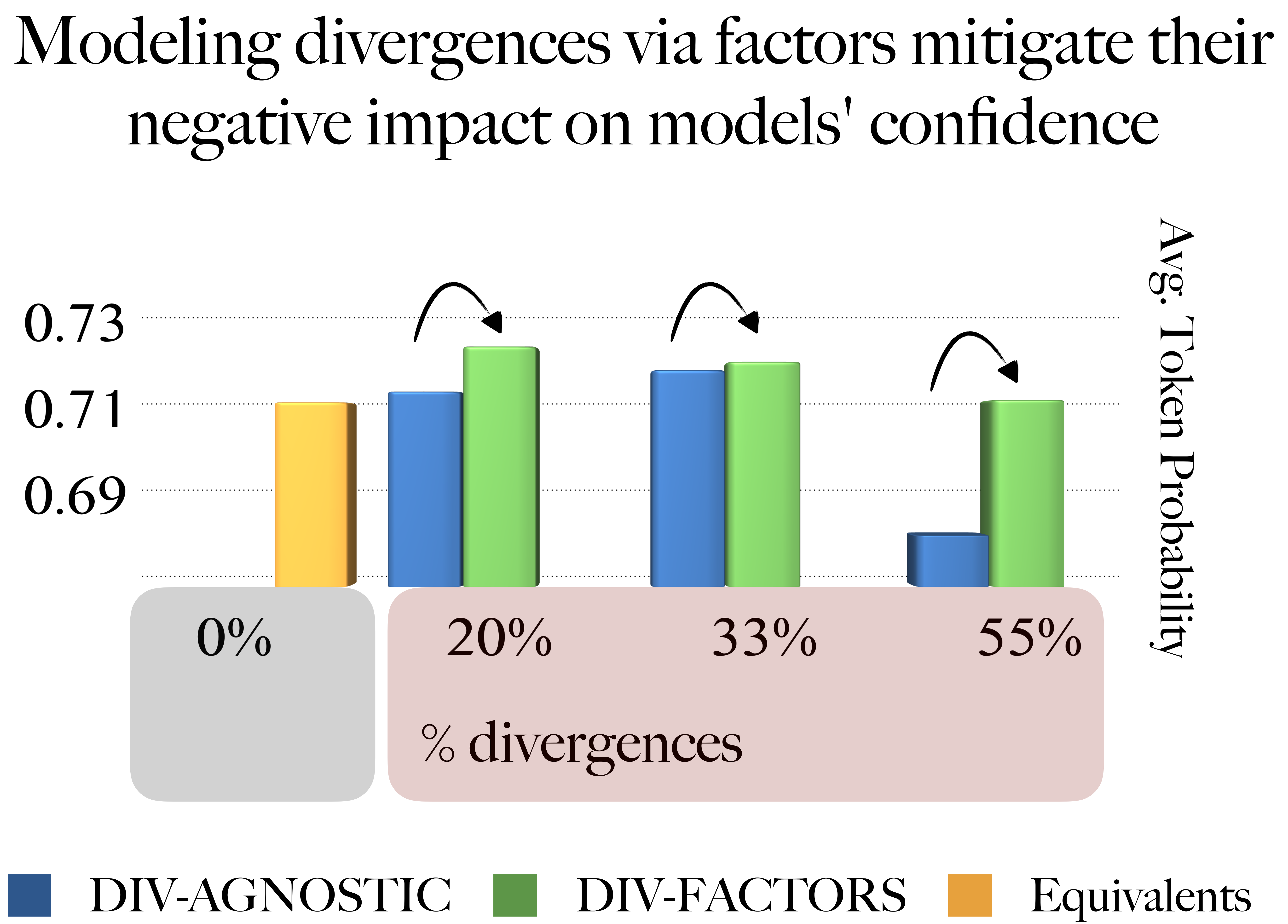
Semantic Divergences decrease translation quality



Modeling divergences via factors help NMT recover from BLEU degradations

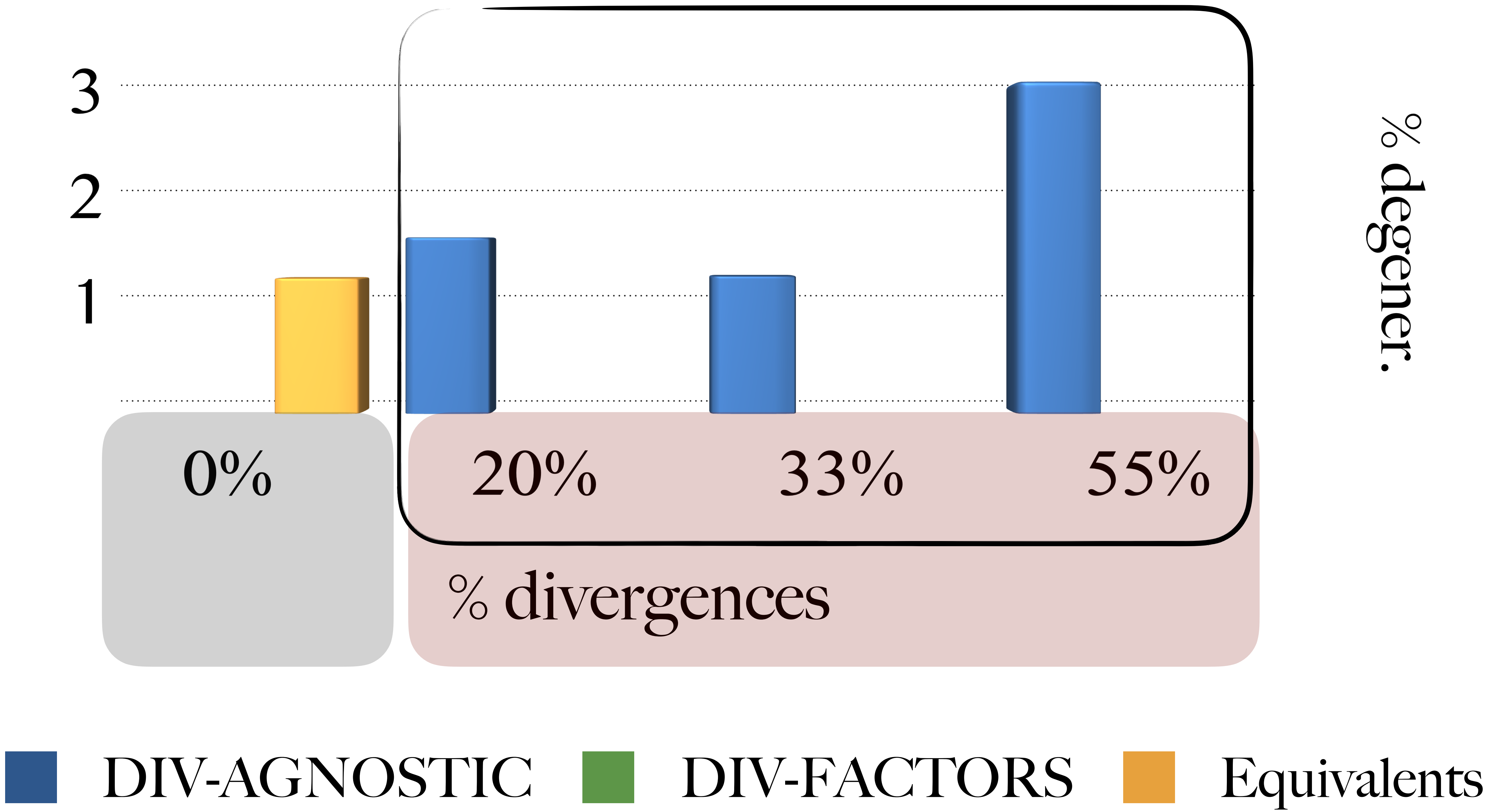



Semantic Divergences decrease
the confidence of token predictions
0.72
0.703
0.687
Avg. Token Probability
0%
20%
33%
55%
% divergences
DIV-AGNOSTIC
DIV-FACTORS
Equivalents

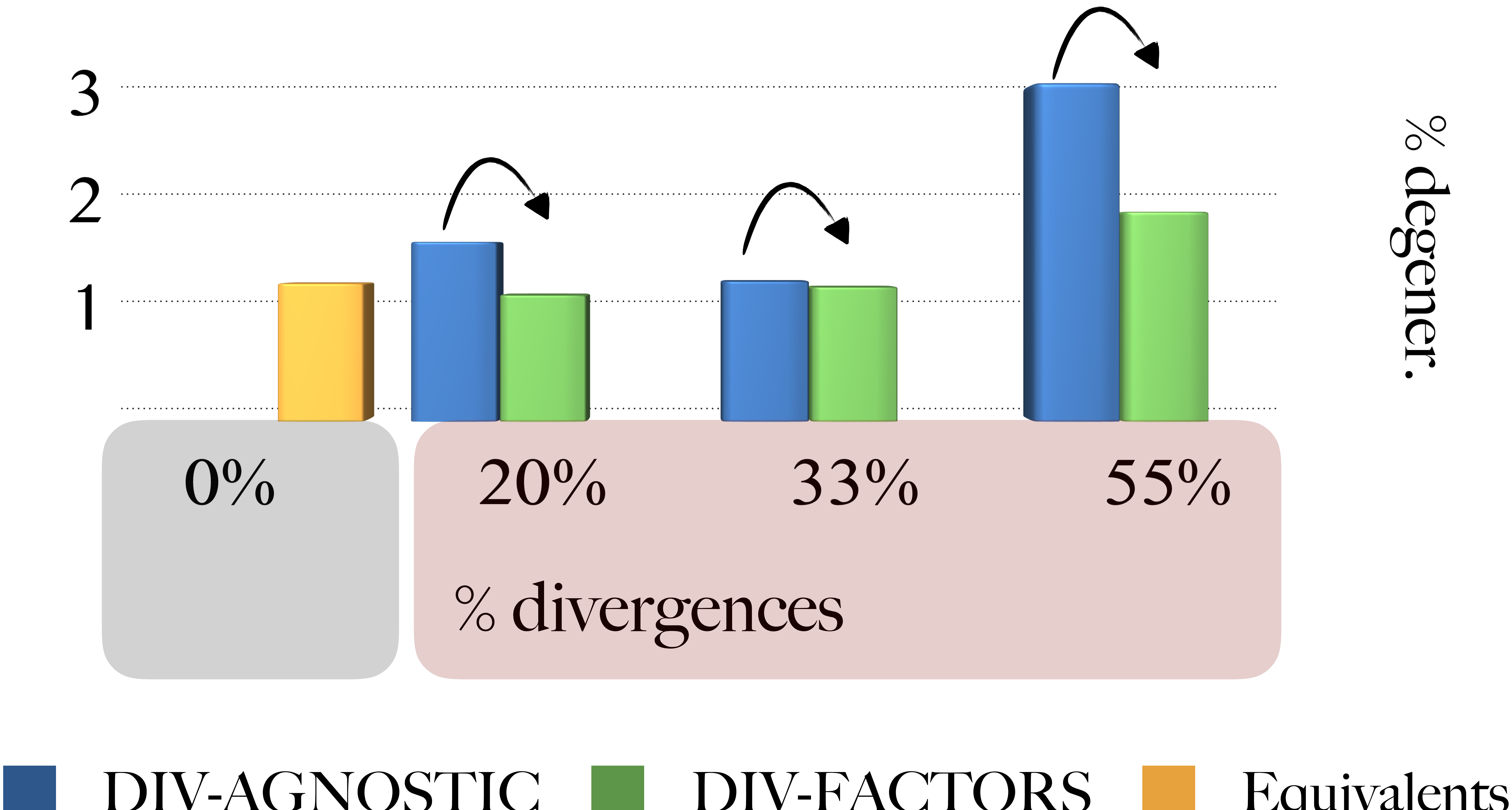


Modeling divergences via factors mitigate their negative impact on models' confidence
0.73
0.71
0.69
Avg. Token Probability
0%
20%
33%
55%
% divergences
DIV-AGNOSTIC
DIV-FACTORS
Equivalents

Semantic Divergences increase the frequency of degenerations



Modeling semantic divergences via factors yield fewer degenerations



Take-aways: Fine-grained distinctions…

🔍 impact various aspects of NMT when they overwhelm the training data

- 👎 hurt translation quality
- 👎 more repetitive loops
- 👎 increase prediction uncertainty

Take-aways: Fine-grained distinctions...

🔍 impact various aspects of NMT when they overwhelm the training data

- 👎 hurt translation quality
- 👎 more repetitive loops
- 👎 increase prediction uncertainty

✎ can inform NMT training

- 👍 encode divergences as token factors
- 👍 mitigate their negative impact

Take-aways: Fine-grained distinctions…

🔎 impact various aspects of NMT when they overwhelm the training data

- 👎 hurt translation quality
- 👎 more repetitive loops
- 👎 increase prediction uncertainty

✎ can inform NMT training

- 👍 encode divergences as token factors
- 👍 mitigate their negative impact

https://github.com/Elbria/xling-SemDiv-NMT