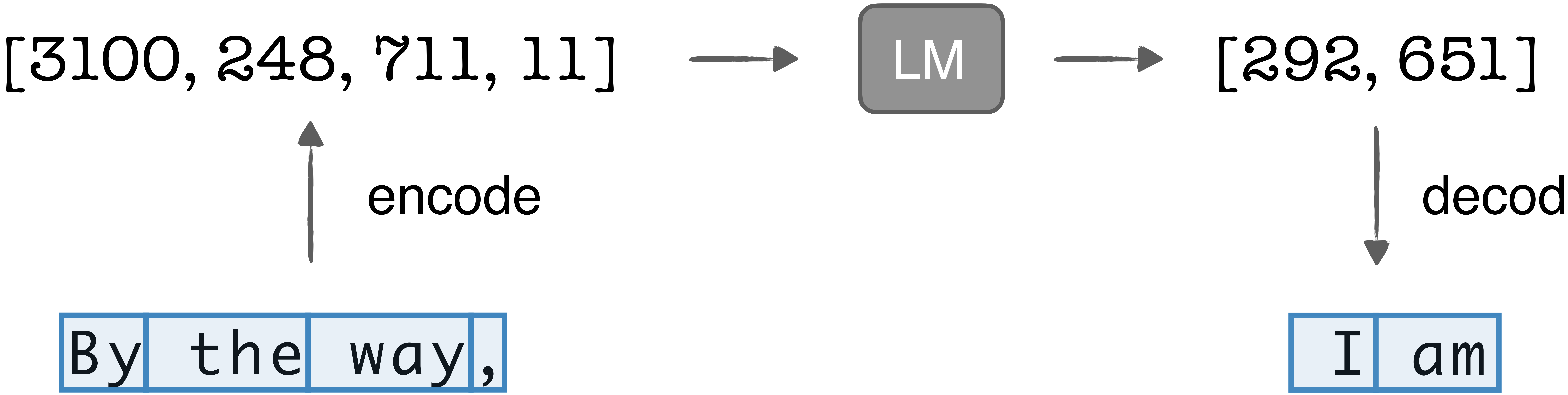
Superword tokenizer for LLMs & Multilingual diversity in vision-text representations

SYNTHDATA @ ICLR 2025

Sewoong Oh, University of Washington

SuperBPE:
Superword Tokenization for Language Models

Tokens are sequences of characters used by LMs to understand text



Modern transformer-based LMs use subword tokenization

- Character-level:

`By the way, I am a fan of the Milky way.`

Modern transformer-based LMs use subword tokenization

- Character-level:

B	y		t	h	e		w	a	y	,		I		a	m		a		f	a	n		o	f		t	h	e		M	i	l	k	y		w	a	y	.

- Word-level:

By	the	way	,	I	am	a	fan	of	the	<UNK>	Way	.

Modern transformer-based LMs use subword tokenization

- Character-level:

  By the way, I am a fan of the Milky way.

- Word-level:

  By the way, I am a fan of the <UNK> Way.

- Subword-level:

  By the way, I am a fan of the Milky Way.

Why do we need to limit tokens to parts of words?

- Multi-word expressions

  "*by the way,*" "*by accident,*" "*for a living,*" "*in the long run*"

- Some languages (e.g., Chinese) do not use **whitespace** at all!

  "*This is a really long sentence that goes on and on*" → "这是一个很长的句子，没完没了"

Byte Pair Encoding (BPE)

Training Data

```
Proof of the Milky Way
consisting of many stars
came in 1610 when Galileo
Galilei used a telescope to
study the Milky Way and
discovered that it is
composed of a huge number
of faint stars.
```

Given **training data** D

Training Data

{Proof, _of, _the, _Milky, _Way, _consisting, _of, _many, _stars, _came, _in, _1610, _when, _Galileo, _Galilei, _used, _a, _telescope, _to, _study, _the, _Milky, _Way, _and, _discovered, _that, _it, _is, _composed, _of, _a, _huge, _number, _of, _faint, _stars.}

Pretokenize D by splitting on whitespace

Training Data

```
_ P r o o f , _ o f , _ t h
e , _ M i l k y , _ W a y , _
c o n s i s t i n g , _ o f ,
_ m a n y , _ s t a r s , _ c
a m e , _ i n , _ 1 6 1 0 , _
w h e n , _ G a l i l e o , _
G a l i l e i , _ u s e d , _
a , _ t e l e s c o p e , _ t
o , _ s t u d y , _ t h e , _
M i l k y , _ W a y , _ a n
d , _ d i s c o v e r e d , _
t h a t , _ i t , _ i s , _ c
o m p o s e d , _ o f , _ a ,
_ h u g e , _ n u m b e r , _
o f , _ f a i n t , _ s t a r
s .
```

Split D into sequence of **bytes**

Training Data

_ P r o o f , _ o f , _ t h
e , _ M i l k y , _ W a y , _
c o n s i s t i n g , _ o f ,
_ m a n y , _ s t a r s , _ c
a m e , _ i n , _ 1 6 1 0 , _
w h e n , _ G a l i l e o , _
G a l i l e i , _ u s e d , _
a , _ t e l e s c o p e , _ t
o , _ s t u d y , _ t h e , _
M i l k y , _ W a y , _ a n
d , _ d i s c o v e r e d , _
t h a t , _ i t , _ i s , _ c
o m p o s e d , _ o f , _ a ,
_ h u g e , _ n u m b e r , _
o f , _ f a i n t , _ s t a r
s .

Pair counts

$\boxed{\_\ t}$	$\boxed{12335282}$
t h	10067390
_ a	9319062
h e	8771183
i n	8024060
e r	6517430
a n	6315205
r e	6031043
o n	5261131
_ i	5209828

Vocabulary

Training Data

_ P r o o f, _ o f, _ t h
e, _ M i l k y, _ W a y, _
c o n s i s t i n g, _ o f,
_ m a n y, _ s t a r s, _ c
a m e, _ i n, _ 1 6 1 0, _
w h e n, _ G a l i l e o, _
G a l i l e i, _ u s e d, _
a, _ t e l e s c o p e, _ t
o, _ s t u d y, _ t h e, _
M i l k y, _ W a y, _ a n
d, _ d i s c o v e r e d, _
t h a t, _ i t, _ i s, _ c
o m p o s e d, _ o f, _ a,
_ h u g e, _ n u m b e r, _
o f, _ f a i n t, _ s t a r
s .

Pair counts

_ t	12335282
t h	10067390
_ a	9319062
h e	8771183
i n	8024060
e r	6517430
a n	6315205
r e	6031043
o n	5261131
_ i	5209828

Vocabulary

_t

Training Data

```
_ P r o o f , _ o f , _t h e ,
_ M i l k y , _ W a y , _ c o
n s i s t i n g , _ o f , _ m
a n y , _ s t a r s , _ c a m
e , _ i n , _ 1 6 1 0 , _ w h
e n , _ G a l i l e o , _ G a
l i l e i , _ u s e d , _ a ,
_t e l e s c o p e , _t o , _
s t u d y , _t h e , _ M i l
k y , _ W a y , _ a n d , _ d
i s c o v e r e d , _t h a
t , _ i t , _ i s , _ c o m p
o s e d , _ o f , _ a , _ h u
g e , _ n u m b e r , _ o f ,
_ f a i n t , _ s t a r s .
```

Pair counts

_ t	12335282
t h	10067390
_ a	9319062
h e	8771183
i n	8024060
e r	6517430
a n	6315205
r e	6031043
o n	5261131
_ i	5209828

Vocabulary

_t

Training Data

_ P r o o f , _ o f , _t h e , _ M i l k y , _ W a y , _ c o n s i s t i n g , _ o f , _ m a n y , _ s t a r s , _ c a m e , _ i n , _ 1 6 1 0 , _ w h e n , _ G a l i l e o , _ G a l i l e i , _ u s e d , _ a , _t e l e s c o p e , _t o , _ s t u d y , _t h e , _ M i l k y , _ W a y , _ a n d , _ d i s c o v e r e d , _t h a t , _ i t , _ i s , _ c o m p o s e d , _ o f , _ a , _ h u g e , _ n u m b e r , _ o f , _ f a i n t , _ s t a r s .

Pair counts

_ a	9319062
h e	8771183
i n	8024060
e r	6517430
a n	6315205
r e	6031043
o n	5261131
_ i	5209828

Vocabulary

_t

Training Data

_ P r o o f , _ o f , _t h e ,
_ M i l k y , _ W a y , _ c o
n s i s t i n g , _ o f , _ m
a n y , _ s t a r s , _ c a m
e , _ i n , _ 1 6 1 0 , _ w h
e n , _ G a l i l e o , _ G a
l i l e i , _ u s e d , _ a ,
_t e l e s c o p e , _t o , _
s t u d y , _t h e , _ M i l
k y , _ W a y , _ a n d , _ d
i s c o v e r e d , _t h a
t , _ i t , _ i s , _ c o m p
o s e d , _ o f , _ a , _ h u
g e , _ n u m b e r , _ o f ,
_ f a i n t , _ s t a r s .

Pair counts

_ a	9319062
h e	8771183
i n	8024060
_t h	7897058
e r	6517430
a n	6315205
r e	6031043
o n	5261131
_ i	5209828

Vocabulary

_t

Training Data

_ P r o o f, _ o f, _t h e,
_ M i l k y, _ W a y, _ c o
n s i s t i n g, _ o f, _ m
a n y, _ s t a r s, _ c a m
e, _ i n, _ 1 6 1 0, _ w h
e n, _ G a l i l e o, _ G a
l i l e i, _ u s e d, _ a,
_t e l e s c o p e, _t o, _
s t u d y, _t h e, _ M i l
k y, _ W a y, _ a n d, _ d
i s c o v e r e d, _t h a
t, _ i t, _ i s, _ c o m p
o s e d, _ o f, _ a, _ h u
g e, _ n u m b e r, _ o f,
_ f a i n t, _ s t a r s .

Pair counts

_ a	9319062
h e	8771183
i n	8024060
_t h	7897058
e r	6517430
a n	6315205
r e	6031043
o n	5261131
_ i	5209828
_ o	5163783

Vocabulary

_t

Training Data

_ P r o o f , _ o f , _t h e ,
_ M i l k y , _ W a y , _ c o
n s i s t i n g , _ o f , _ m
a n y , _ s t a r s , _ c a m
e , _ i n , _ 1 6 1 0 , _ w h
e n , _ G a l i l e o , _ G a
l i l e i , _ u s e d , _ a ,
_t e l e s c o p e , _t o , _
s t u d y , _t h e , _ M i l
k y , _ W a y , _ a n d , _ d
i s c o v e r e d , _t h a
t , _ i t , _ i s , _ c o m p
o s e d , _ o f , _ a , _ h u
g e , _ n u m b e r , _ o f ,
_ f a i n t , _ s t a r s .

Pair counts

_ a	9319062
h e	8771183
i n	8024060
_t h	7897058
e r	6517430
a n	6315205
r e	6031043
o n	5261131
_ i	5209828
_ o	5163783

Vocabulary

_t

Training Data

```
_ P r o o f , _ o f , _t h e ,
_ M i l k y , _ W a y , _ c o
n s i s t i n g , _ o f , _ m
a n y , _ s t a r s , _ c a m
e , _ i n , _ 1 6 1 0 , _ w h
e n , _ G a l i l e o , _ G a
l i l e i , _ u s e d , _ a ,
_t e l e s c o p e , _t o , _
s t u d y , _t h e , _ M i l
k y , _ W a y , _ a n d , _ d
i s c o v e r e d , _t h a
t , _ i t , _ i s , _ c o m p
o s e d , _ o f , _ a , _ h u
g e , _ n u m b e r , _ o f ,
_ f a i n t , _ s t a r s .
```

Pair counts

_ a	9319062
h e	8771183
i n	8024060
_t h	7897058
e r	6517430
a n	6315205
r e	6031043
o n	5261131
_ i	5209828
_ o	5163783

Vocabulary

_t

_a

Training Data

_ P r o o f , _ o f , _t h e ,
_ M i l k y , _ W a y , _ c o
n s i s t i n g , _ o f , _ m
a n y , _ s t a r s , _ c a m
e , _ i n , _ 1 6 1 0 , _ w h
e n , _ G a l i l e o , _ G a
l i l e i , _ u s e d , _a ,
_t e l e s c o p e , _t o , _
s t u d y , _t h e , _ M i l
k y , _ W a y , _a n d , _ d i
s c o v e r e d , _t h a t ,
_ i t , _ i s , _ c o m p o s
e d , _ o f , _a , _ h u g e ,
_ n u m b e r , _ o f , _ f a
i n t , _ s t a r s .

Pair counts

_ a	9319062
h e	8771183
i n	8024060
_t h	7897058
e r	6517430
a n	6315205
r e	6031043
o n	5261131
_ i	5209828
_ o	5163783

Vocabulary

_t
_a

Training Data

_ P r o o f, _ o f, _t h e, _ M i l k y, _ W a y, _ c o n s i s t i n g, _ o f, _ m a n y, _ s t a r s, _ c a m e, _ i n, _ 1 6 1 0, _ w h e n, _ G a l i l e o, _ G a l i l e i, _ u s e d, _a, _t e l e s c o p e, _t o, _ s t u d y, _t h e, _ M i l k y, _ W a y, _a n d, _ d i s c o v e r e d, _t h a t, _ i t, _ i s, _ c o m p o s e d, _ o f, _a, _ h u g e, _ n u m b e r, _ o f, _ f a i n t, _ s t a r s .

Pair counts

h e	8771183
i n	8024060
_t h	7897058
e r	6517430
r e	6031043
o n	5261131
_ i	5209828
_ o	5163783

Vocabulary

_t

_a

Training Data

_ P r o o f , _ o f , _t h e ,
_ M i l k y , _ W a y , _ c o
n s i s t i n g , _ o f , _ m
a n y , _ s t a r s , _ c a m
e , _ i n , _ 1 6 1 0 , _ w h
e n , _ G a l i l e o , _ G a
l i l e i , _ u s e d , _a ,
_t e l e s c o p e , _t o , _
s t u d y , _t h e , _ M i l
k y , _ W a y , _a n d , _ d i
s c o v e r e d , _t h a t ,
_ i t , _ i s , _ c o m p o s
e d , _ o f , _a , _ h u g e ,
_ n u m b e r , _ o f , _ f a
i n t , _ s t a r s .

Pair counts

h e	8771183
i n	8024060
_t h	7897058
e r	6517430
r e	6031043
o n	5261131
_ i	5209828
_ o	5163783
_ s	5035505
_ w	4523998

Vocabulary

_t
_a

Training Data

_ P r o o f , _ o f , _t h e ,
_ M i l k y , _ W a y , _ c o
n s i s t i n g , _ o f , _ m
a n y , _ s t a r s , _ c a m
e , _ i n , _ 1 6 1 0 , _ w h
e n , _ G a l i l e o , _ G a
l i l e i , _ u s e d , _a ,
_t e l e s c o p e , _t o , _
s t u d y , _t h e , _ M i l
k y , _ W a y , _a n d , _ d i
s c o v e r e d , _t h a t ,
_ i t , _ i s , _ c o m p o s
e d , _ o f , _a , _ h u g e ,
_ n u m b e r , _ o f , _ f a
i n t , _ s t a r s .

Pair counts

h e	8771183
i n	8024060
_t h	7897058
e r	6517430
r e	6031043
o n	5261131
_ i	5209828
_ o	5163783
_ s	5035505
_ w	4523998

Vocabulary

_t
_a

Training Data

_ P r o o f, _ o f, _t h e, _ M i l k y, _ W a y, _ c o n s i s t i n g, _ o f, _ m a n y, _ s t a r s, _ c a m e, _ i n, _ 1 6 1 0, _ w h e n, _ G a l i l e o, _ G a l i l e i, _ u s e d, _a, _t e l e s c o p e, _t o, _ s t u d y, _t h e, _ M i l k y, _ W a y, _a n d, _ d i s c o v e r e d, _t h a t, _ i t, _ i s, _ c o m p o s e d, _ o f, _a, _ h u g e, _ n u m b e r, _ o f, _ f a i n t, _ s t a r s .

Pair counts

h e	8771183
i n	8024060
_t h	7897058
e r	6517430
r e	6031043
o n	5261131
_ i	5209828
_ o	5163783
_ s	5035505
_ w	4523998

Vocabulary

_t
_a
he

Training Data

_ P r o o f, _ o f, _t he,
_ M i l k y, _ W a y, _ c o
n s i s t i n g, _ o f, _ m
a n y, _ s t a r s, _ c a m
e, _ i n, _ 1 6 1 0, _ w he
n, _ G a l i l e o, _ G a l
i l e i, _ u s e d, _a, _t
e l e s c o p e, _t o, _ s
t u d y, _t he, _ M i l k
y, _ W a y, _a n d, _ d i s
c o v e r e d, _t h a t, _
i t, _ i s, _ c o m p o s e
d, _ o f, _a, _ h u g e, _
n u m b e r, _ o f, _ f a i
n t, _ s t a r s .

Pair counts

h e	8771183
i n	8024060
_t h	7897058
e r	6517430
r e	6031043
o n	5261131
_ i	5209828
_ o	5163783
_ s	5035505
_ w	4523998

Vocabulary

_t
_a
he

Training Data

_ P r o o f, _ o f, _t he,
_ M i l k y, _ W a y, _ c o
n s i s t i n g, _ o f, _ m
a n y, _ s t a r s, _ c a m
e, _ i n, _ 1 6 1 0, _ w he
n, _ G a l i l e o, _ G a l
i l e i, _ u s e d, _a, _t
e l e s c o p e, _t o, _ s
t u d y, _t he, _ M i l k
y, _ W a y, _a n d, _ d i s
c o v e r e d, _t h a t, _
i t, _ i s, _ c o m p o s e
d, _ o f, _a, _ h u g e, _
n u m b e r, _ o f, _ f a i
n t, _ s t a r s .

Pair counts

i n	8024060
e r	6517430
r e	6031043
o n	5261131
_ i	5209828
_ o	5163783
_ s	5035505
_ w	4523998

Vocabulary

_t
_a
he

Training Data

_ P r o o f , _ o f , _t he ,
_ M i l k y , _ W a y , _ c o
n s i s t i n g , _ o f , _ m
a n y , _ s t a r s , _ c a m
e , _ i n , _ 1 6 1 0 , _ w he
n , _ G a l i l e o , _ G a l
i l e i , _ u s e d , _a , _t
e l e s c o p e , _t o , _ s
t u d y , _t he , _ M i l k
y , _ W a y , _a n d , _ d i s
c o v e r e d , _t h a t , _
i t , _ i s , _ c o m p o s e
d , _ o f , _a , _ h u g e , _
n u m b e r , _ o f , _ f a i
n t , _ s t a r s .

Pair counts

i n	8024060
r e	6031043
_t he	5605612
e r	5279258
o n	5261131
_ i	5209828
_ o	5163783
_ s	5035505
_ w	4523998

Vocabulary

_t
_a
he

Training Data

_ P r o o f, _ o f, _t he,
_ M i l k y, _ W a y, _ c o
n s i s t i n g, _ o f, _ m
a n y, _ s t a r s, _ c a m
e, _ i n, _ 1 6 1 0, _ w he
n, _ G a l i l e o, _ G a l
i l e i, _ u s e d, _a, _t
e l e s c o p e, _t o, _ s
t u d y, _t he, _ M i l k
y, _ W a y, _a n d, _ d i s
c o v e r e d, _t h a t, _
i t, _ i s, _ c o m p o s e
d, _ o f, _a, _ h u g e, _
n u m b e r, _ o f, _ f a i
n t, _ s t a r s .

Pair counts

i n	8024060
r e	6031043
_t he	5605612
e r	5279258
o n	5261131
_ i	5209828
_ o	5163783
_ s	5035505
_ w	4523998
a t	4424733

Vocabulary

_t
_a
he

Training Data

```
_ P r o o f, _ o f, _t he,
_ M i l k y, _ W a y, _ c o
n s i s t i n g, _ o f, _ m
a n y, _ s t a r s, _ c a m
e, _ i n, _ 1 6 1 0, _ w he
n, _ G a l i l e o, _ G a l
i l e i, _ u s e d, _a, _t
e l e s c o p e, _t o, _ s
t u d y, _t he, _ M i l k
y, _ W a y, _a n d, _ d i s
c o v e r e d, _t h a t, _
i t, _ i s, _ c o m p o s e
d, _ o f, _a, _ h u g e, _
n u m b e r, _ o f, _ f a i
n t, _ s t a r s .
```

Pair counts

i n	8024060
r e	6031043
_t he	5605612
e r	5279258
o n	5261131
_ i	5209828
_ o	5163783
_ s	5035505
_ w	4523998
a t	4424733

Vocabulary

_t

_a

he

Training Data

```
_ P r o o f , _ o f , _t he ,
_ M i l k y , _ W a y , _ c o
n s i s t i n g , _ o f , _ m
a n y , _ s t a r s , _ c a m
e , _ i n , _ 1 6 1 0 , _ w he
n , _ G a l i l e o , _ G a l
i l e i , _ u s e d , _a , _t
e l e s c o p e , _t o , _ s
t u d y , _t he , _ M i l k
y , _ W a y , _a n d , _ d i s
c o v e r e d , _t h a t , _
i t , _ i s , _ c o m p o s e
d , _ o f , _a , _ h u g e , _
n u m b e r , _ o f , _ f a i
n t , _ s t a r s .
```

Pair counts

i n	8024060
r e	6031043
_t he	5605612
e r	5279258
o n	5261131
_ i	5209828
_ o	5163783
_ s	5035505
_ w	4523998
a t	4424733

Vocabulary

_t

_a

he

in

Training Data

```
_ P r o o f , _ o f , _t he ,
_ M i l k y , _ W a y , _ c o
n s i s t in g , _ o f , _ m
a n y , _ s t a r s , _ c a m
e , _ in , _ 1 6 1 0 , _ w he
n , _ G a l i l e o , _ G a l
i l e i , _ u s e d , _a , _t
e l e s c o p e , _t o , _ s
t u d y , _t he , _ M i l k
y , _ W a y , _a n d , _ d i s
c o v e r e d , _t h a t , _
i t , _ i s , _ c o m p o s e
d , _ o f , _a , _ h u g e , _
n u m b e r , _ o f , _ f a
in t , _ s t a r s .
```

Pair counts

i n	8024060
r e	6031043
_t he	5605612
e r	5279258
o n	5261131
_ i	5209828
_ o	5163783
_ s	5035505
_ w	4523998
a t	4424733

Vocabulary

_t

_a

he

in

Training Data

```
_ P r o o f, _ o f, _t he,
_ M i l k y, _ W a y, _ c o
n s i s t in g, _ o f, _ m
a n y, _ s t a r s, _ c a m
e, _ in, _ 1 6 1 0, _ w he
n, _ G a l i l e o, _ G a l
i l e i, _ u s e d, _a, _t
e l e s c o p e, _t o, _ s
t u d y, _t he, _ M i l k
y, _ W a y, _a n d, _ d i s
c o v e r e d, _t h a t, _
i t, _ i s, _ c o m p o s e
d, _ o f, _a, _ h u g e, _
n u m b e r, _ o f, _ f a
in t, _ s t a r s .
```

Pair counts

r e	6031043
_t he	5605612
e r	5279258
o n	5261131
_ o	5163783
_ s	5035505
_ w	4523998
a t	4424733
o r	4162447
e s	4010515

Vocabulary

_t

_a

he

in

Training Data

_ P r o o f , _ o f , _t he ,
_ M i l k y , _ W a y , _ c o
n s i s t in g , _ o f , _ m
a n y , _ s t a r s , _ c a m
e , _ in , _ 1 6 1 0 , _ w he
n , _ G a l i l e o , _ G a l
i l e i , _ u s e d , _a , _t
e l e s c o p e , _t o , _ s
t u d y , _t he , _ M i l k
y , _ W a y , _a n d , _ d i s
c o v e r e d , _t h a t , _
i t , _ i s , _ c o m p o s e
d , _ o f , _a , _ h u g e , _
n u m b e r , _ o f , _ f a
in t , _ s t a r s .

Pair counts

r e	6031043
_t he	5605612
e r	5279258
o n	5261131
_ o	5163783
_ s	5035505
_ w	4523998
a t	4424733
o r	4162447
e s	4010515

Vocabulary

_t
_a
he
in

Training Data

```
_ P r o o f, _ o f, _t he,
_ M i l k y, _ W a y, _ c o
n s i s t in g, _ o f, _ m
a n y, _ s t a r s, _ c a m
e, _ in, _ 1 6 1 0, _ w he
n, _ G a l i l e o, _ G a l
i l e i, _ u s e d, _a, _t
e l e s c o p e, _t o, _ s
t u d y, _t he, _ M i l k
y, _ W a y, _a n d, _ d i s
c o v e r e d, _t h a t, _
i t, _ i s, _ c o m p o s e
d, _ o f, _a, _ h u g e, _
n u m b e r, _ o f, _ f a
in t, _ s t a r s .
```

Pair counts

r e	6031043
_t he	5605612
e r	5279258
o n	5261131
_ o	5163783
_ s	5035505
_ w	4523998
a t	4424733
o r	4162447
e s	4010515

Vocabulary

- _t
- _a
- he
- in
- re

Training Data

```
_ P r o o f, _ o f, _t he,
_ M i l k y, _ W a y, _ c o
n s i s t in g, _ o f, _ m
a n y, _ s t a r s, _ c a m
e, _ in, _ 1 6 1 0, _ w he
n, _ G a l i l e o, _ G a l
i l e i, _ u s e d, _a, _t
e l e s c o p e, _t o, _ s
t u d y, _t he, _ M i l k
y, _ W a y, _a n d, _ d i s
c o v e re d, _t h a t, _ i
t, _ i s, _ c o m p o s e
d, _ o f, _a, _ h u g e, _
n u m b e r, _ o f, _ f a
in t, _ s t a r s .
```

Pair counts

r e	6031043
_t he	5605612
e r	5279258
o n	5261131
_ o	5163783
_ s	5035505
_ w	4523998
a t	4424733
o r	4162447
e s	4010515

Vocabulary

_t
_a
he
in
re

Training Data

```
_ P r o o f, _ o f, _t he,
_ M i l k y, _ W a y, _ c o
n s i s t in g, _ o f, _ m
a n y, _ s t a r s, _ c a m
e, _ in, _ 1 6 1 0, _ w he
n, _ G a l i l e o, _ G a l
i l e i, _ u s e d, _a, _t
e l e s c o p e, _t o, _ s
t u d y, _t he, _ M i l k
y, _ W a y, _a n d, _ d i s
c o v e re d, _t h a t, _ i
t, _ i s, _ c o m p o s e
d, _ o f, _a, _ h u g e, _
n u m b e r, _ o f, _ f a
in t, _ s t a r s .
```

Pair counts

_t he	5605612
o n	5261131
_ o	5163783
_ s	5035505
e r	4754849
_ w	4523998
a t	4424733
o u	3838417
_ c	3831635
n d	3811435

Vocabulary

_t
_a
he
in
re

Training Data

_ P r o o f , _ o f , _t he ,
_ M i l k y , _ W a y , _ c o
n s i s t in g , _ o f , _ m
a n y , _ s t a r s , _ c a m
e , _ in , _ 1 6 1 0 , _ w he
n , _ G a l i l e o , _ G a l
i l e i , _ u s e d , _a , _t
e l e s c o p e , _t o , _ s
t u d y , _t he , _ M i l k
y , _ W a y , _a n d , _ d i s
c o v e re d , _t h a t , _ i
t , _ i s , _ c o m p o s e
d , _ o f , _a , _ h u g e , _
n u m b e r , _ o f , _ f a
in t , _ s t a r s .

Pair counts

_t he	5605612
o n	5261131
_ o	5163783
_ s	5035505
e r	4754849
_ w	4523998
a t	4424733
o u	3838417
_ c	3831635
n d	3811435

Vocabulary

_t
_a
he
in
re

Training Data

_ P r o o f, _ o f, _t he,
_ M i l k y, _ W a y, _ c o
n s i s t in g, _ o f, _ m
a n y, _ s t a r s, _ c a m
e, _ in, _ 1 6 1 0, _ w he
n, _ G a l i l e o, _ G a l
i l e i, _ u s e d, _a, _t
e l e s c o p e, _t o, _ s
t u d y, _t he, _ M i l k
y, _ W a y, _a n d, _ d i s
c o v e re d, _t h a t, _ i
t, _ i s, _ c o m p o s e
d, _ o f, _a, _ h u g e, _
n u m b e r, _ o f, _ f a
in t, _ s t a r s .

Pair counts

_t he	5605612
o n	5261131
_ o	5163783
_ s	5035505
e r	4754849
_ w	4523998
a t	4424733
o u	3838417
_ c	3831635
n d	3811435

Vocabulary

_t
_a
he
in
re
_the

Training Data

_ P r o o f, _ o f, _the, _
M i l k y, _ W a y, _ c o n
s i s t in g, _ o f, _ m a
n y, _ s t a r s, _ c a m
e, _ in, _ 1 6 1 0, _ w he
n, _ G a l i l e o, _ G a l
i l e i, _ u s e d, _a, _t
e l e s c o p e, _t o, _ s
t u d y, _the, _ M i l k y,
_ W a y, _a n d, _ d i s c
o v e re d, _t h a t, _ i
t, _ i s, _ c o m p o s e
d, _ o f, _a, _ h u g e, _
n u m b e r, _ o f, _ f a
in t, _ s t a r s .

Pair counts

_t he	5605612
o n	5261131
_ o	5163783
_ s	5035505
e r	4754849
_ w	4523998
a t	4424733
o u	3838417
_ c	3831635
n d	3811435

Vocabulary

_t
_a
he
in
re
_the

Training Data

```
_ P r o o f, _ o f, _the, _
M i l k y, _ W a y, _ c o n
s i s t in g, _ o f, _ m a
n y, _ s t a r s, _ c a m
e, _ in, _ 1 6 1 0, _ w he
n, _ G a l i l e o, _ G a l
i l e i, _ u s e d, _a, _t
e l e s c o p e, _t o, _ s
t u d y, _the, _ M i l k y,
_ W a y, _a n d, _ d i s c
o v e re d, _t h a t, _ i
t, _ i s, _ c o m p o s e
d, _ o f, _a, _ h u g e, _
n u m b e r, _ o f, _ f a
in t, _ s t a r s .
```

Pair counts

o n	5261131
_ o	5163783
_ s	5035505
e r	4754849
_ w	4523998
a t	4424733
o u	3838417
_ c	3831635
n d	3811435
o r	3661288

Vocabulary

_t
_a
he
in
re
_the

Training Data

_ P r o o f, _ o f, _the, _
M i l k y, _ W a y, _ c o n
s i s t in g, _ o f, _ m a
n y, _ s t a r s, _ c a m
e, _ in, _ 1 6 1 0, _ w he
n, _ G a l i l e o, _ G a l
i l e i, _ u s e d, _a, _t
e l e s c o p e, _t o, _ s
t u d y, _the, _ M i l k y,
_ W a y, _a n d, _ d i s c
o v e re d, _t h a t, _ i
t, _ i s, _ c o m p o s e
d, _ o f, _a, _ h u g e, _
n u m b e r, _ o f, _ f a
in t, _ s t a r s .

Pair counts

o n	5261131
_ o	5163783
_ s	5035505
e r	4754849
_ w	4523998
a t	4424733
o u	3838417
_ c	3831635
n d	3811435
o r	3661288

Vocabulary

_t
_a
he
in
re
_the
⋮

until we reach desired vocab size T

SuperBPE

- Phase 1: Run BPE with whitespace barrier from pretokenization until $t<T$
- Phase 2: Run BPE without whitespace barrier until T
- Intuition: learn the basic units of meaning (words) in the first phase, and then merge common word sequences (superwords)

SuperBPE

- Phase 1: Run BPE with whitespace barrier from pretokenization until t<T
- Phase 2: Run BPE without whitespace barrier until T
- Intuition: learn the basic units of meaning (words) in the first phase, and then merge common word sequences (superwords)

POS tag	#	Random examples
NN, IN	906	_case_of, _depend_on, _availability_of, _emphasis_on, _distinction_between
VB, DT	566	_reached_a, _discovered_the, _identify_the, _becomes_a, _issued_a
DT, NN	498	_this_month, _no_idea, _the_earth, _the_maximum, _this_stuff
IN, NN	406	_on_top, _by_accident, _in_effect, _for_lunch, _in_front
IN, DT, NN	333	_for_a_living, _by_the_way, _into_the_future, _in_the_midst
IN, DT, NN, IN	33	_at_the_time_of, _in_the_presence_of, _in_the_middle_of, _in_a_way_that

Training Data

Proof of the Milky Way consisting of many stars came in 1610 when Galileo Galilei used a telescope to study the Milky Way and discovered that it is composed of a huge number of faint stars.

Training Data

```
{Proof_of_the_Milky_Way_co
nsisting_of_many_stars_cam
e_in_, 1610,
_when_Galileo_Galilei_used
_a_telescope_to_study_the_
Milky_Way_and_discovered_t
hat_it_is_composed_of_a_hu
ge_number_of_faint_stars.}
```

- 2nd phase:
  - Skip whitespace pretokenization
  - but can still use other pretokenization rules, e.g., numbers

Training Data

```
{P r o o f _ o f _ t h e _
M i l k y _ W a y _ c o n s
i s t i n g _ o f _ m a n y
_ s t a r s _ c a m e _ i
n, _ 1 6 1 0, _ w h e n _ G
a l i l e o _ G a l i l e i
_ u s e d _ a _ t e l e s c
o p e _ t o _ s t u d y _ t
h e _ M i l k y _ W a y _ a
n d _ d i s c o v e r e d _
t h a t _ i t _ i s _ c o m
p o s e d _ o f _ a _ h u g
e _ n u m b e r _ o f _ f a
i n t _ s t a r s .}
```

Split D into sequence of bytes

Training Data

{Proof _of _the _Milky _Way _consisting _of _many _stars _came _in_, 1 610, _when _Gal ileo _Galilei _used _a _telescope _to _study _the _Milky _Way _and _discovered _that _it _is _composed _of _a _huge _number _of _faint _stars.}

Apply tokenizer learned so far

Training Data

{Proof _of _the _Milky _Way
_consisting _of _many
_stars _came _in_, 1 610,
_when _Gal ileo _Galilei
_used _a _telescope _to
_study _the _Milky _Way
_and _discovered _that _it
_is _composed _of _a _huge
_number _of _faint _stars.}

Pair counts

_of _the	517482
' s	456028
, _and	413189
_in _the	362529
' t	247975
. _The	232178
, _the	226412
_to _the	222524
, _but	200360
_on _the	164233

Vocabulary

stage 1

_t
_a
he
in
re
_the
⋮
_Aleg

Training Data

{Proof _of _the _Milky _Way _consisting _of _many _stars _came _in_, 1 610, _when _Gal ileo _Galilei _used _a _telescope _to _study _the _Milky _Way _and _discovered _that _it _is _composed _of _a _huge _number _of _faint _stars.}

Pair counts

_of _the	517482
' s	456028
, _and	413189
_in _the	362529
' t	247975
. _The	232178
, _the	226412
_to _the	222524
, _but	200360
_on _the	164233

Vocabulary

stage 1

_t
_a
he
in
re
_the
⋮
_Aleg

_of _the

Training Data

{Proof _of_the _Milky _Way
_consisting _of _many
_stars _came _in_, 1 610,
_when _Gal ileo _Galilei
_used _a _telescope _to
_study _the _Milky _Way
_and _discovered _that _it
_is _composed _of _a _huge
_number _of _faint _stars.}

Pair counts

_of	_the	517482
'	s	456028
,	_and	413189
_in	_the	362529
'	t	247975
.	_The	232178
,	_the	226412
_to	_the	222524
,	_but	200360
_on	_the	164233

Vocabulary

stage 1

_t
_a
he
in
re
_the
⋮
_Aleg

_of _the

Training Data

{Proof _of_the _Milky _Way
_consisting _of _many
_stars _came _in_, 1 610,
_when _Gal ileo _Galilei
_used _a _telescope _to
_study _the _Milky _Way
_and _discovered _that _it
_is _composed _of _a _huge
_number _of _faint _stars.}

Pair counts

' s	456028
, _and	413189
_in _the	362529
' t	247975
. _The	232178
, _the	226412
_to _the	222524
, _but	200360
_on _the	164233
. _I	159471

Vocabulary

stage 1

_t
_a
he
in
re
_the
⋮
_Aleg

_of _the

Training Data

{Proof _of_the _Milky _Way
_consisting _of _many
_stars _came _in_, 1 610,
_when _Gal ileo _Galilei
_used _a _telescope _to
_study _the _Milky _Way
_and _discovered _that _it
_is _composed _of _a _huge
_number _of _faint _stars.}

Pair counts

' s	456028
, _and	413189
_in _the	362529
' t	247975
. _The	232178
, _the	226412
_to _the	222524
, _but	200360
_on _the	164233
. _I	159471

Vocabulary

stage 1

_t
_a
he
in
re
_the
⋮
_Aleg

_of _the

Training Data

{Proof _of_the _Milky _Way
_consisting _of _many
_stars _came _in_, 1 610,
_when _Gal ileo _Galilei
_used _a _telescope _to
_study _the _Milky _Way
_and _discovered _that _it
_is _composed _of _a _huge
_number _of _faint _stars.}

Pair counts

' s	456028
, _and	413189
_in _the	362529
' t	247975
. _The	232178
, _the	226412
_to _the	222524
, _but	200360
_on _the	164233
. _I	159471

Vocabulary

stage 1

_t
_a
he
in
re
_the
⋮
_Aleg

_of _the

' s

Training Data

{Proof _of_the _Milky _Way
_consisting _of _many
_stars _came _in_, 1 610,
_when _Gal ileo _Galilei
_used _a _telescope _to
_study _the _Milky _Way
_and _discovered _that _it
_is _composed _of _a _huge
_number _of _faint _stars.}

Pair counts

, _and	413189
_in _the	362529
' t	247975
. _The	232178
, _the	226412
_to _the	222524
, _but	200360
_on _the	164233
. _I	159471
? _	148101

Vocabulary

stage 1

_t
_a
he
in
re
_the
⋮
_Aleg

_of _the

' s

Training Data

{Proof _of_the _Milky _Way _consisting _of _many _stars _came _in_, 1 610, _when _Gal ileo _Galilei _used _a _telescope _to _study _the _Milky _Way _and _discovered _that _it _is _composed _of _a _huge _number _of _faint _stars.}

Pair counts

, _and	413189
_in _the	362529
' t	247975
. _The	232178
, _the	226412
_to _the	222524
, _but	200360
_on _the	164233
. _I	159471
? _	148101

Vocabulary

stage 1

_t
_a
he
in
re
_the
⋮
_Aleg

_of _the

' s

Training Data

{Proof _of_the _Milky _Way
_consisting _of _many
_stars _came _in_, 1 610,
_when _Gal ileo _Galilei
_used _a _telescope _to
_study _the _Milky _Way
_and _discovered _that _it
_is _composed _of _a _huge
_number _of _faint _stars.}

Pair counts

, _and	413189
_in _the	362529
' t	247975
. _The	232178
, _the	226412
_to _the	222524
, _but	200360
_on _the	164233
. _I	159471
? _	148101

Vocabulary

stage 1

_t
_a
he
in
re
_the
⋮
_Aleg

_of _the

' s

, _and

Training Data

{Proof _of_the _Milky _Way
_consisting _of _many
_stars _came _in_, 1 610,
_when _Gal ileo _Galilei
_used _a _telescope _to
_study _the _Milky _Way
_and _discovered _that _it
_is _composed _of _a _huge
_number _of _faint _stars.}

Pair counts

_in _the	362529
' t	247975
. _The	232178
, _the	226412
_to _the	222524
, _but	200360
_on _the	164233
. _I	159471
? _	148101
_to _be	147449

Vocabulary

stage 1

_t
_a
he
in
re
_the
⋮
_Aleg

_of _the

' s

, _and

Training Data

{Proof _of_the _Milky _Way
_consisting _of _many
_stars _came _in_, 1 610,
_when _Gal ileo _Galilei
_used _a _telescope _to
_study _the _Milky _Way
_and _discovered _that _it
_is _composed _of _a _huge
_number _of _faint _stars.}

Pair counts

_in _the	362529
' t	247975
. _The	232178
, _the	226412
_to _the	222524
, _but	200360
_on _the	164233
. _I	159471
? _	148101
_to _be	147449

Vocabulary

stage 1

_t
_a
he
in
re
_the
⋮
_Aleg

_of _the

' s

, _and

Training Data

{Proof _of_the _Milky _Way _consisting _of _many _stars _came _in_, 1 610, _when _Gal ileo _Galilei _used _a _telescope _to _study _the _Milky _Way _and _discovered _that _it _is _composed _of _a _huge _number _of _faint _stars.}

Pair counts

_in _the	362529
' t	247975
. _The	232178
, _the	226412
_to _the	222524
, _but	200360
_on _the	164233
. _I	159471
? _	148101
_to _be	147449

Vocabulary

stage 1

_t
_a
he
in
re
_the
⋮
_Aleg

_of _the

' s

, _and

_in _the

Training Data

{Proof _of_the _Milky _Way
_consisting _of _many
_stars _came _in_, 1 610,
_when _Gal ileo _Galilei
_used _a _telescope _to
_study _the _Milky _Way
_and _discovered _that _it
_is _composed _of _a _huge
_number _of _faint _stars.}

Pair counts

_in _the	362529
' t	247975
. _The	232178
, _the	226412
_to _the	222524
, _but	200360
_on _the	164233
. _I	159471
? _	148101
_to _be	147449

Vocabulary

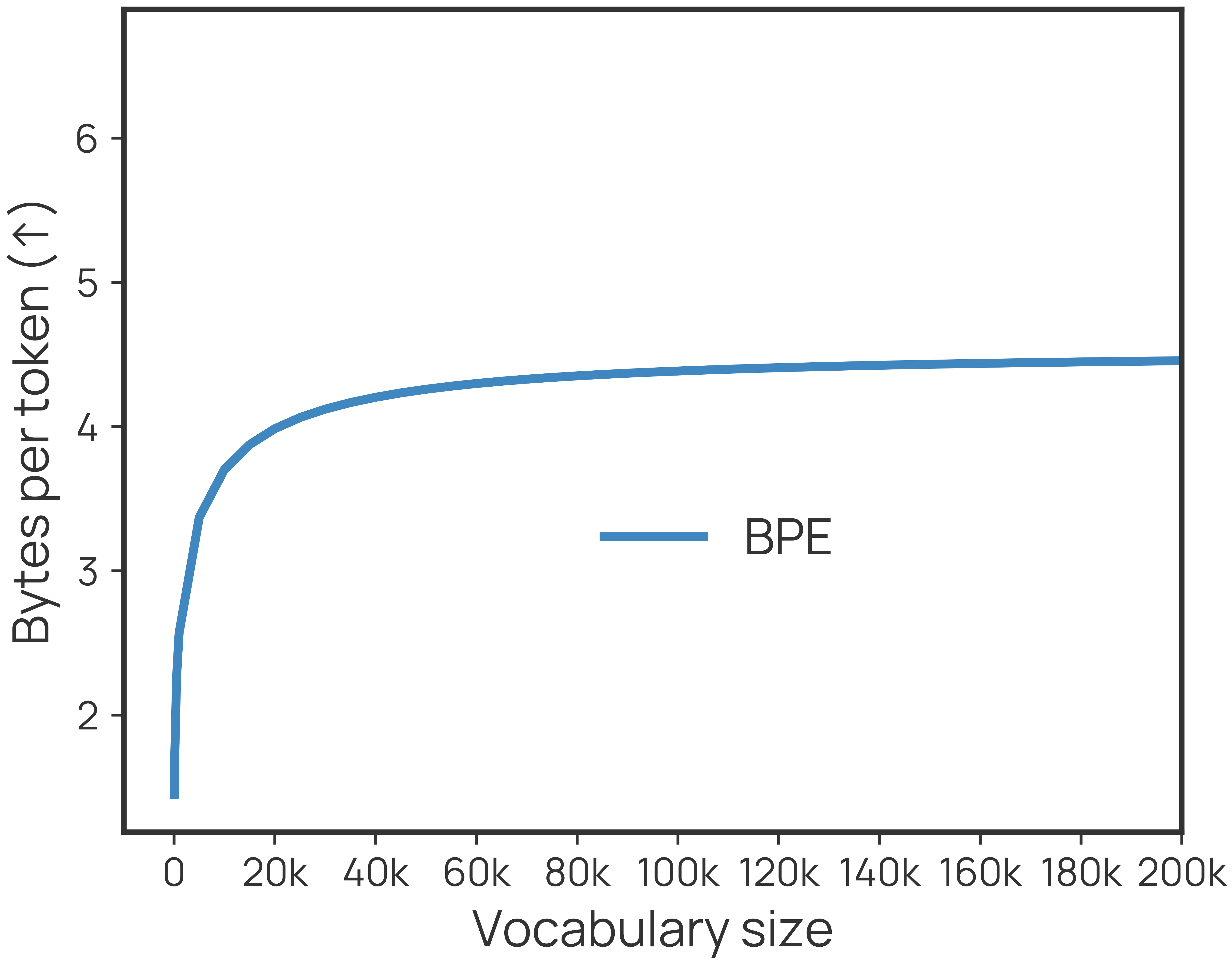
stage 1

_t
_a
he
in
re
_the
⋮
_Aleg

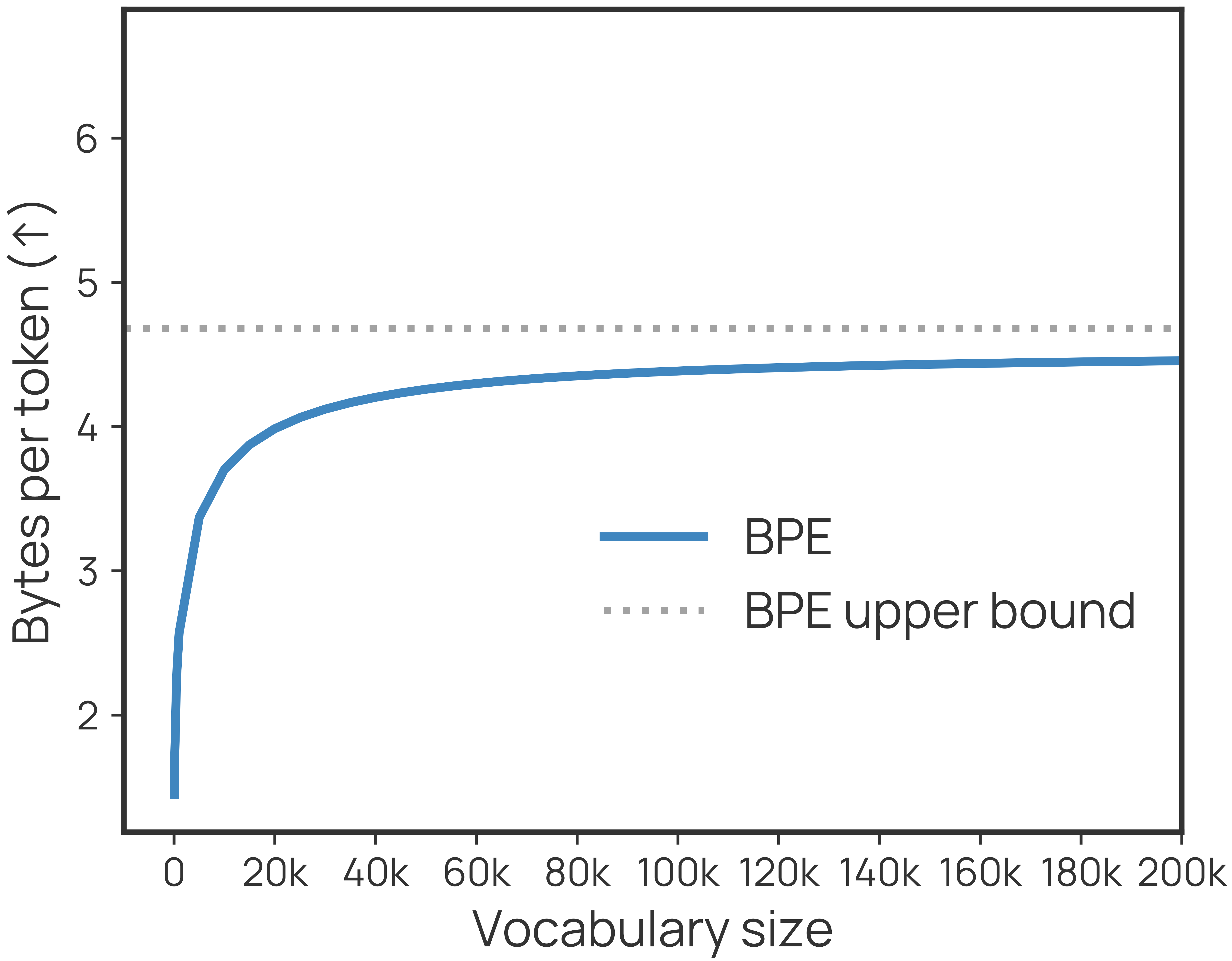
_of _the
' s
, _and
_in _the
⋮

until we reach desired vocab size T

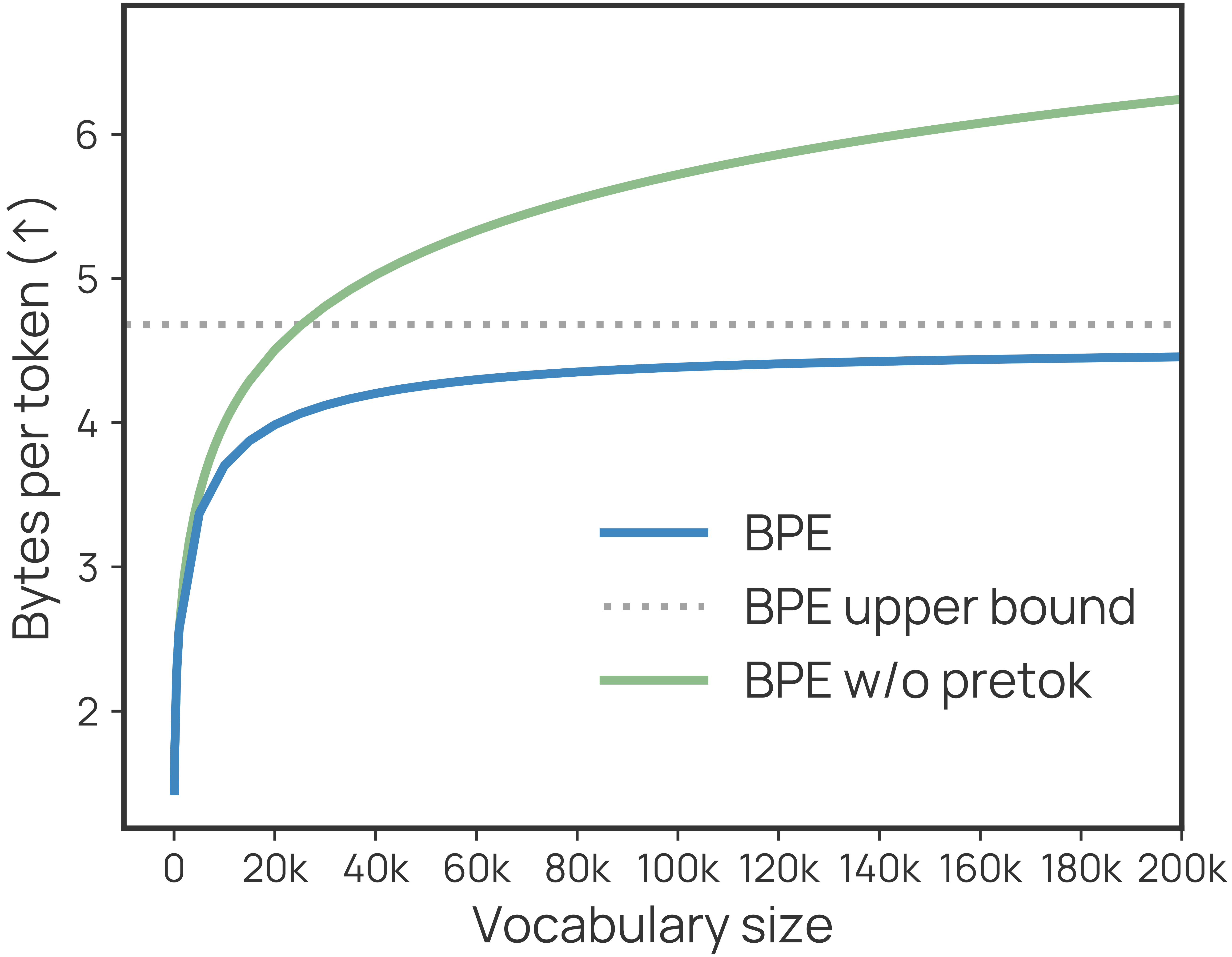
SuperBPE encodes text more efficiently



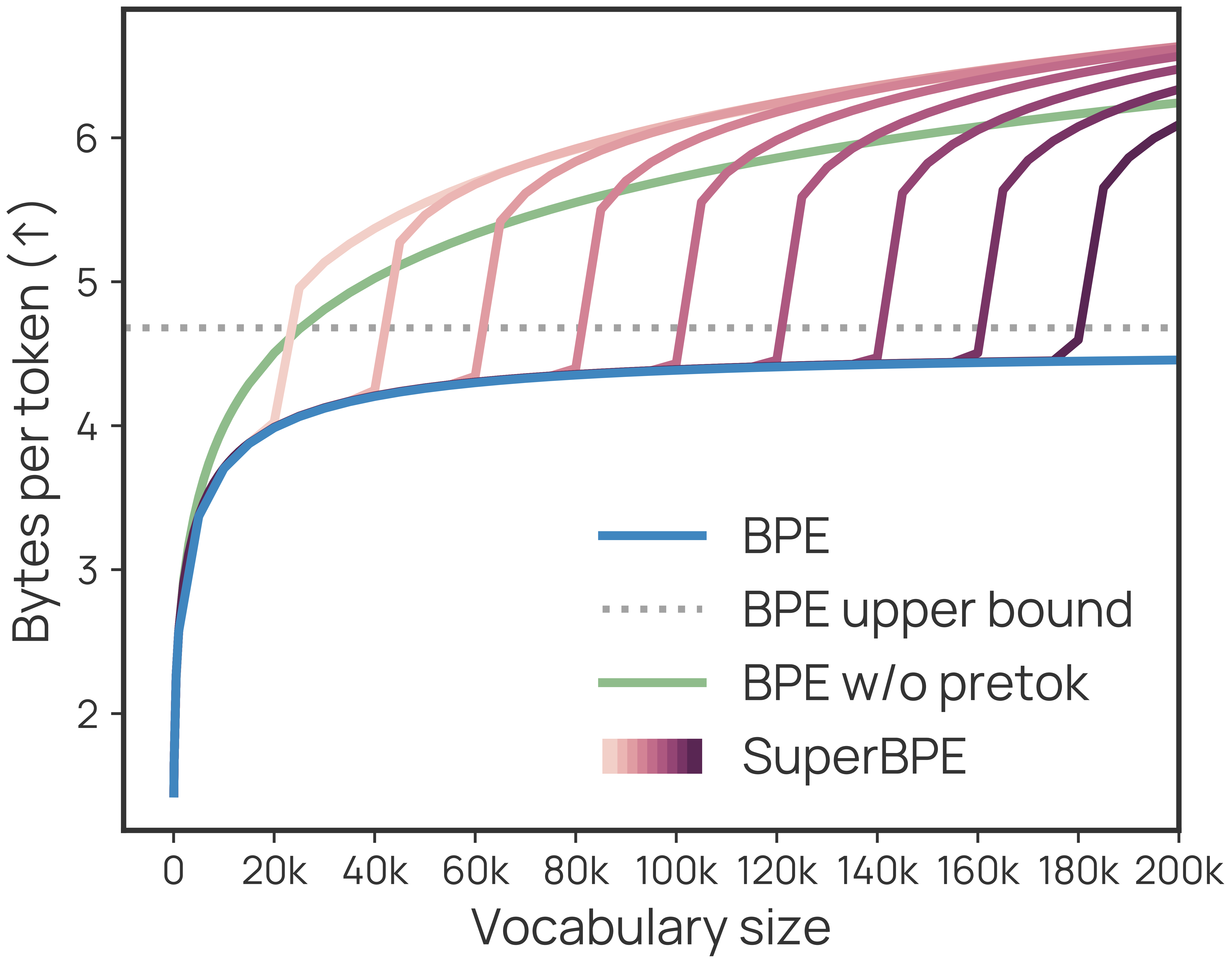
SuperBPE encodes text more efficiently



SuperBPE encodes text more efficiently



SuperBPE encodes text more efficiently

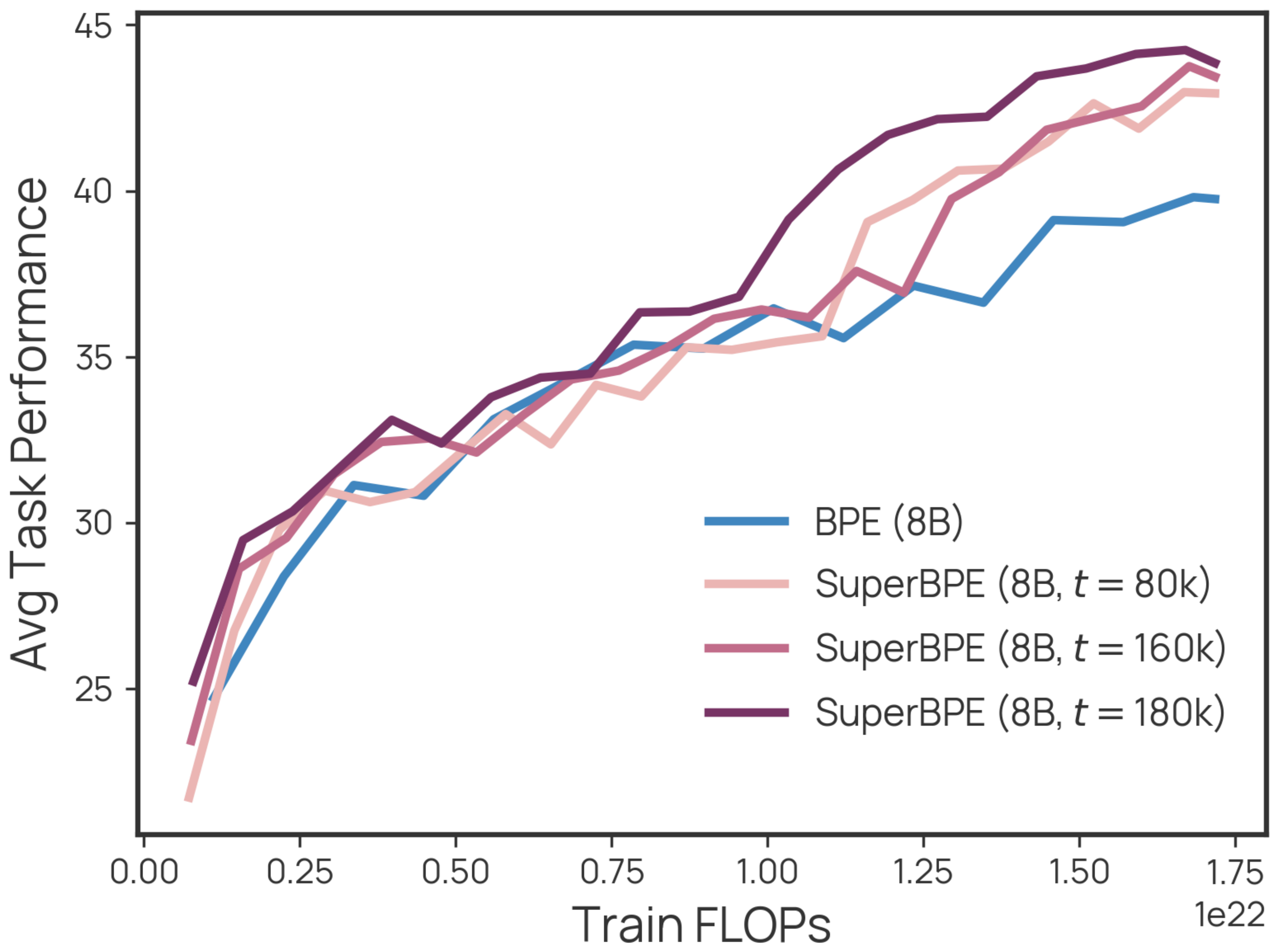


Changing tokenizer requires pretraining LLM

Baseline: **BPE 8B** (Olmo2 @ 330B tokens)

SuperBPE 8B

- ✅ model size
- ✅ training compute
- ❌ inference compute (35% less)
- ❌ amount of text seen (41% more)



Takeaways

- SuperBPE extends subword BPE to let tokens include superwords, or (parts of) multiple words
- SuperBPE needs about 33% less tokens to encode the same context
- Given same amount of compute, we can pretrain on more text to achieve improved downstream performance

Multilingual diversity in vision-text representations

The same text, “stove”, means different things depending on the region/culture

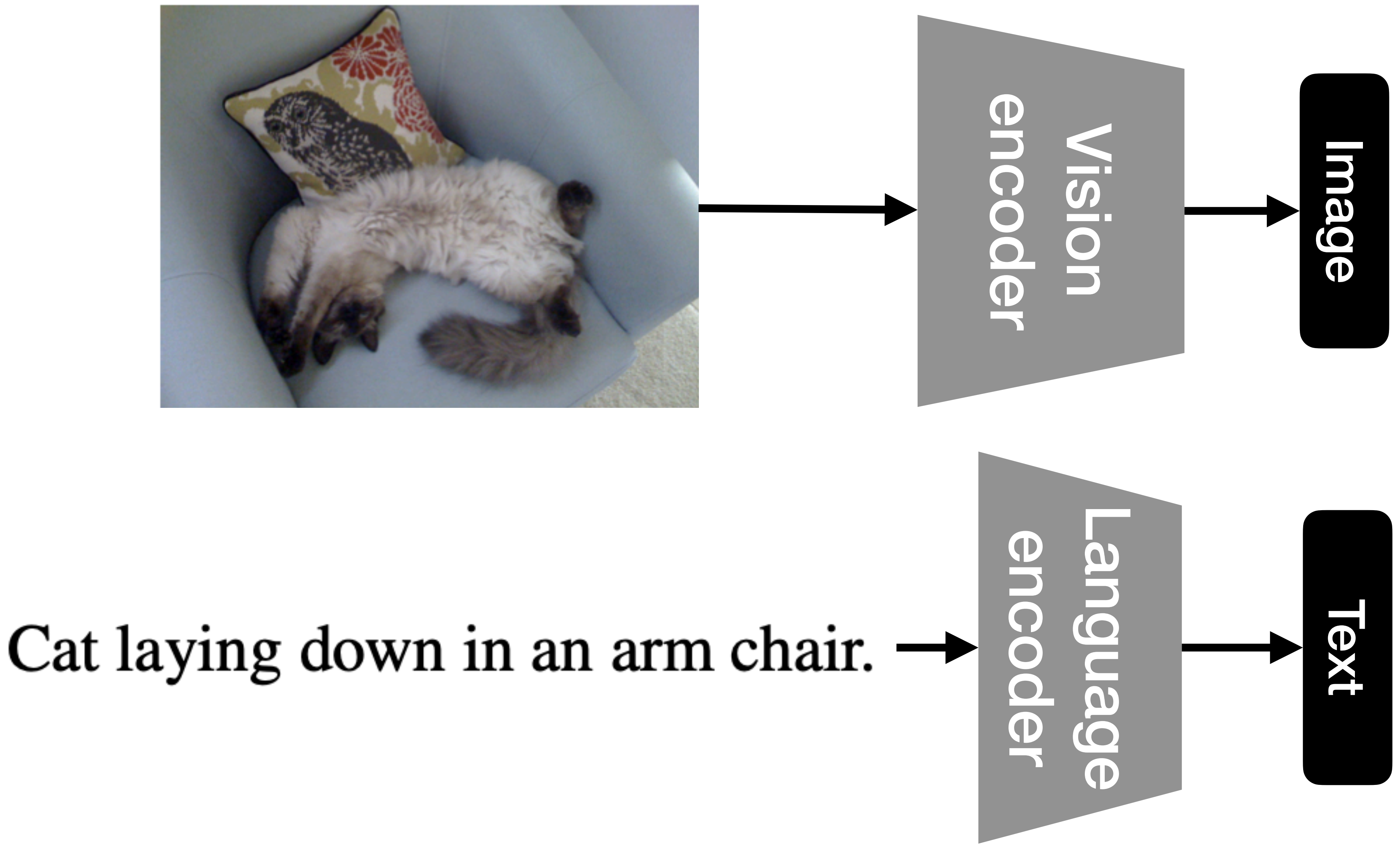
The same image is perceived differently depending on the background

	English	I.	Luggage left unattended at a table.
		II.	Luggage lined up next to tables with jackets resting on the tables.
		III.	luggage sitting next to tables
	Japanese	I.	A man sitting in a lobby with lots of suitcases and bags
		II.	A man is sitting in a room, and there are several tables filled with luggage nearby.
		III.	Man waiting with a lot of luggage
	English	I.	Cat laying down in an arm chair.
		II.	A Siamese cat laying on its back on a couch next to a pillow.
		III.	A cat stretched out and upside down on a chair
	Japanese	I.	A cat stretches out on a blue chair and a pillow with an embroidered owl next to it.
		II.	A cat is relaxing next to a cushion with a picture of an owl on it.
		III.	Cat sitting on his back in an armchair with an owl-patternedcushion

A Ye, S Santy, JD Hwang, AX Zhang, R Krishna, “Semantic and Expressive Variation in Image Captions Across Languages”, CVPR 2025

Benchmark training datasets suffer monolingual biases

- CLIP (Contrastive Language-Image Pre-training) model is trained on the contrastive loss of the inner-product

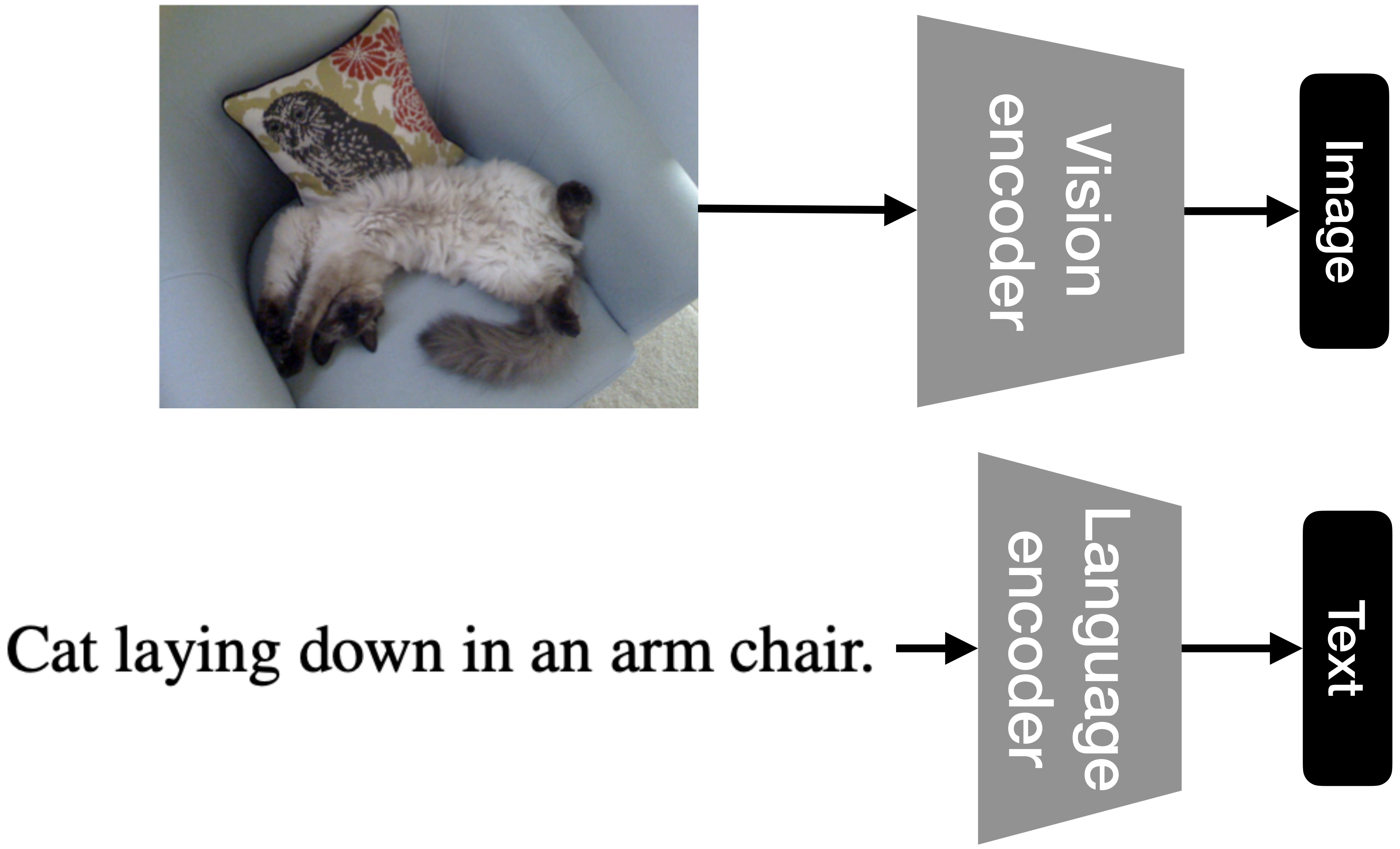


CLIP score: $\langle \text{Image}, \text{Text} \rangle$

Radford et al., "Learning transferable visual models from natural language supervision", ICML 2021

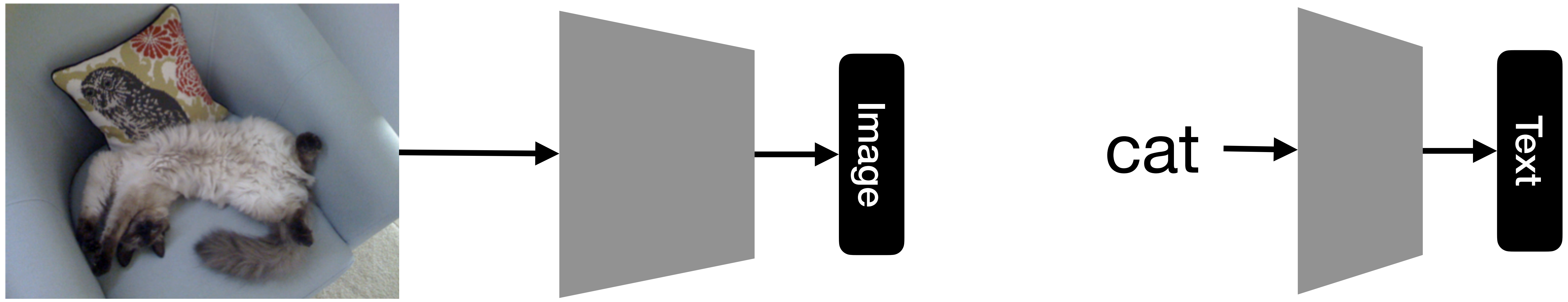
Benchmark training datasets suffer monolingual biases

- CLIP (Contrastive Language-Image Pre-training) model is trained on the contrastive loss of the inner-product



- Zero-shot evaluation on ImageNet, for example

Baselines (1.28B samples seen)	ImageNet
Filtered raw captions	0.414



Radford et al., "Learning transferable visual models from natural language supervision", ICML 2021

Benchmark training datasets suffer monolingual biases

- **LAION** and **DataComp:** open-source CLIP training datasets

1. Internet-scale data pool

( , Cat laying down in an arm chair.),

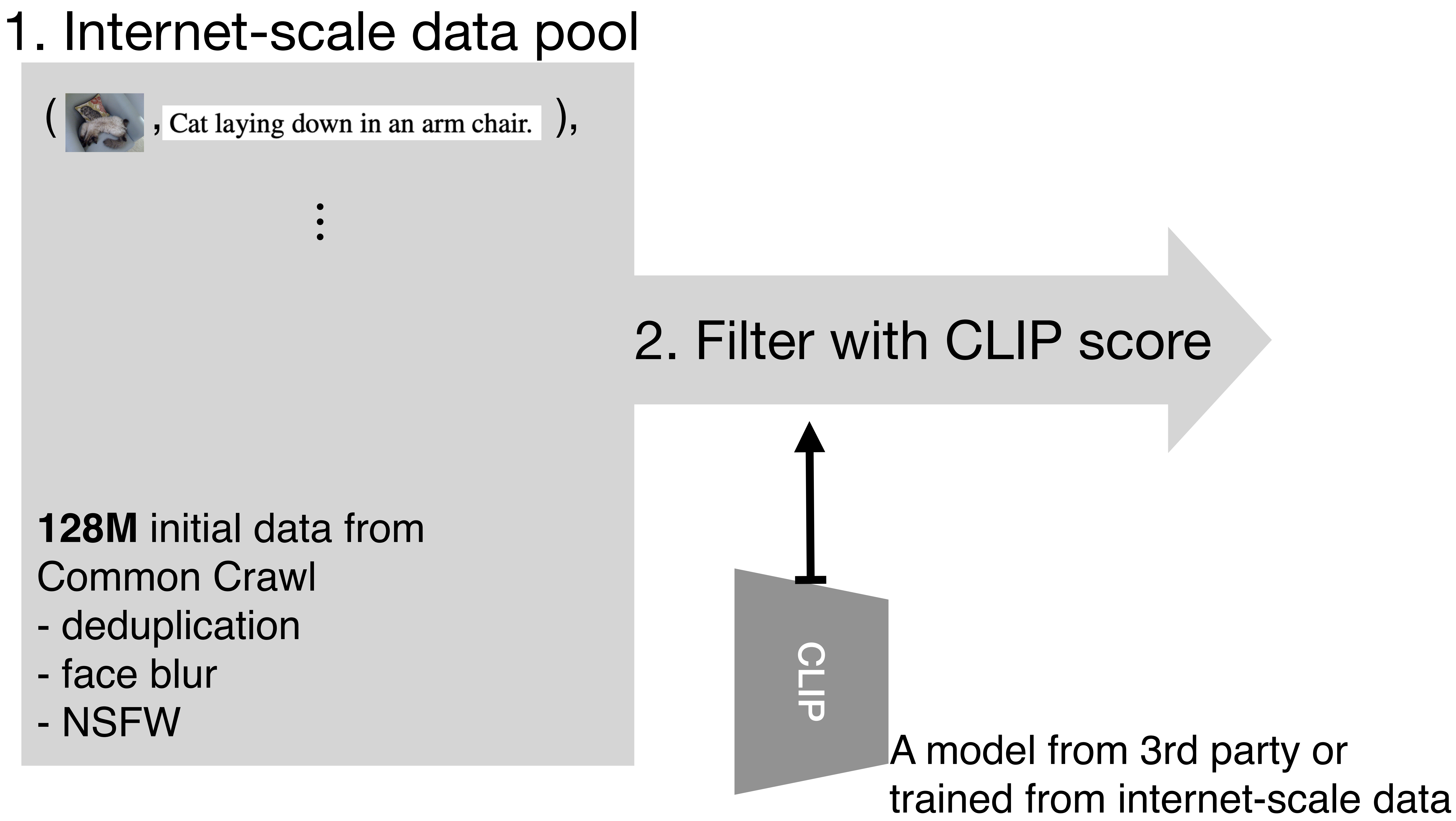
⋮

128M initial data from
Common Crawl
- deduplication
- face blur
- NSFW

Gadre et al. “Datacomp: In search of the next generation of multimodal datasets”, NeurIPS 2023

Benchmark training datasets suffer monolingual biases

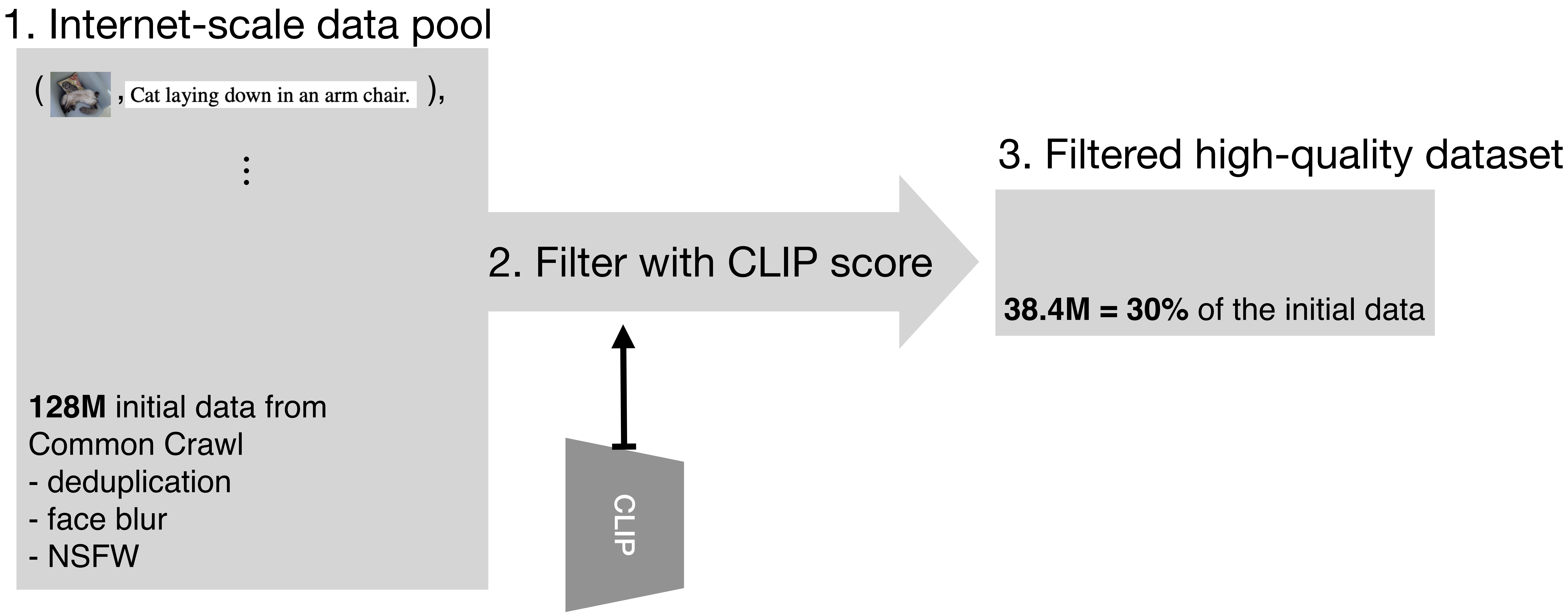
- **LAION** and **DataComp:** open-source CLIP training datasets



Gadre et al. “Datacomp: In search of the next generation of multimodal datasets”, NeurIPS 2023

Benchmark training datasets suffer monolingual biases

- **LAION** and **DataComp:** open-source CLIP training datasets



Gadre et al. "Datacomp: In search of the next generation of multimodal datasets", NeurIPS 2023

Benchmark training datasets suffer monolingual biases

- **LAION** and **DataComp:** open-source CLIP training datasets

1. Internet-scale data pool

( , Cat laying down in an arm chair.),

⋮

128M initial data from Common Crawl
- deduplication
- face blur
- NSFW

2. Filter with CLIP score

CLIP

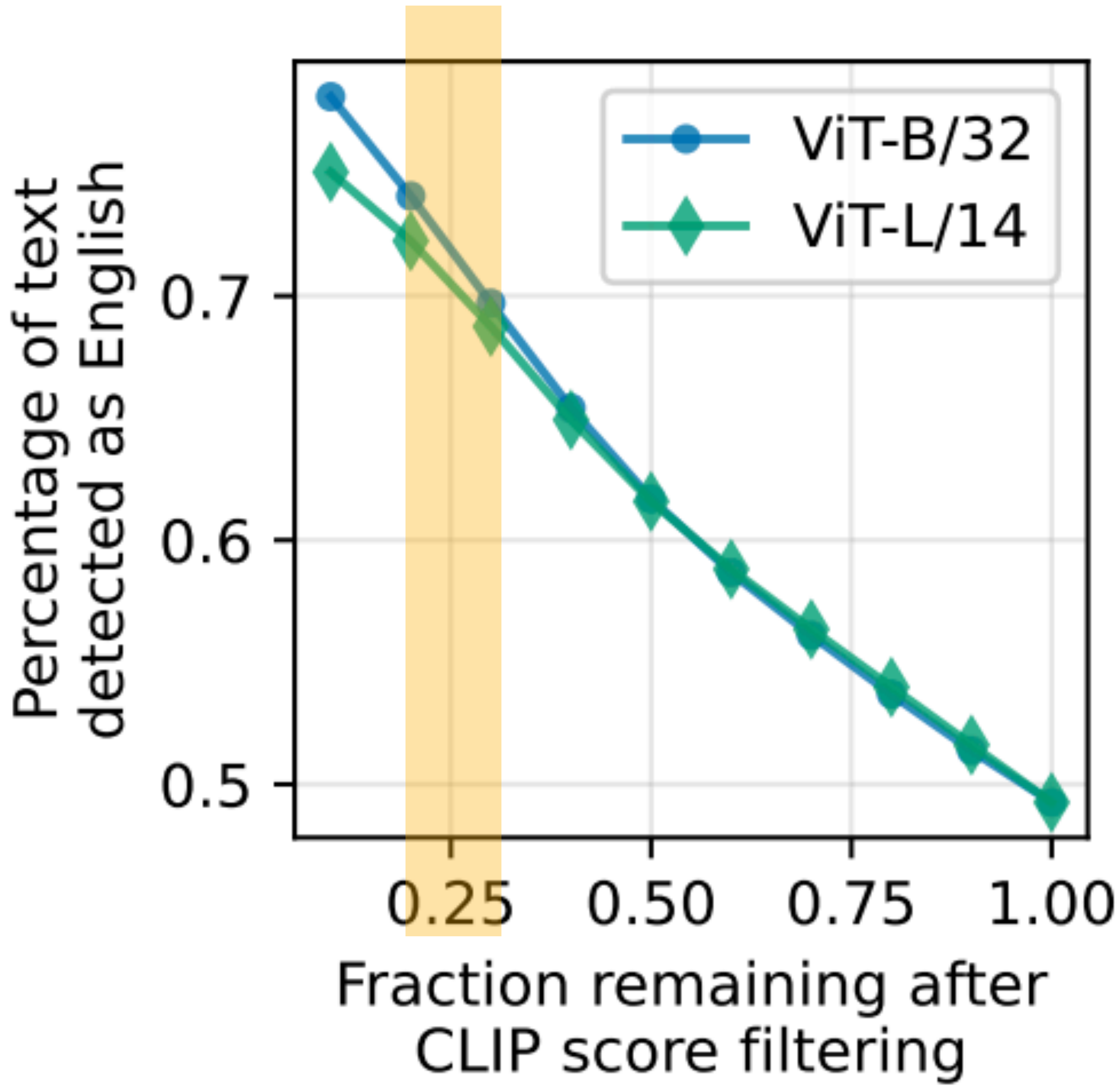
3. Filtered high-quality dataset

38.4M = 30% of the initial data

Baselines (1.28B samples seen)	ImageNet
Filtered raw captions	0.414

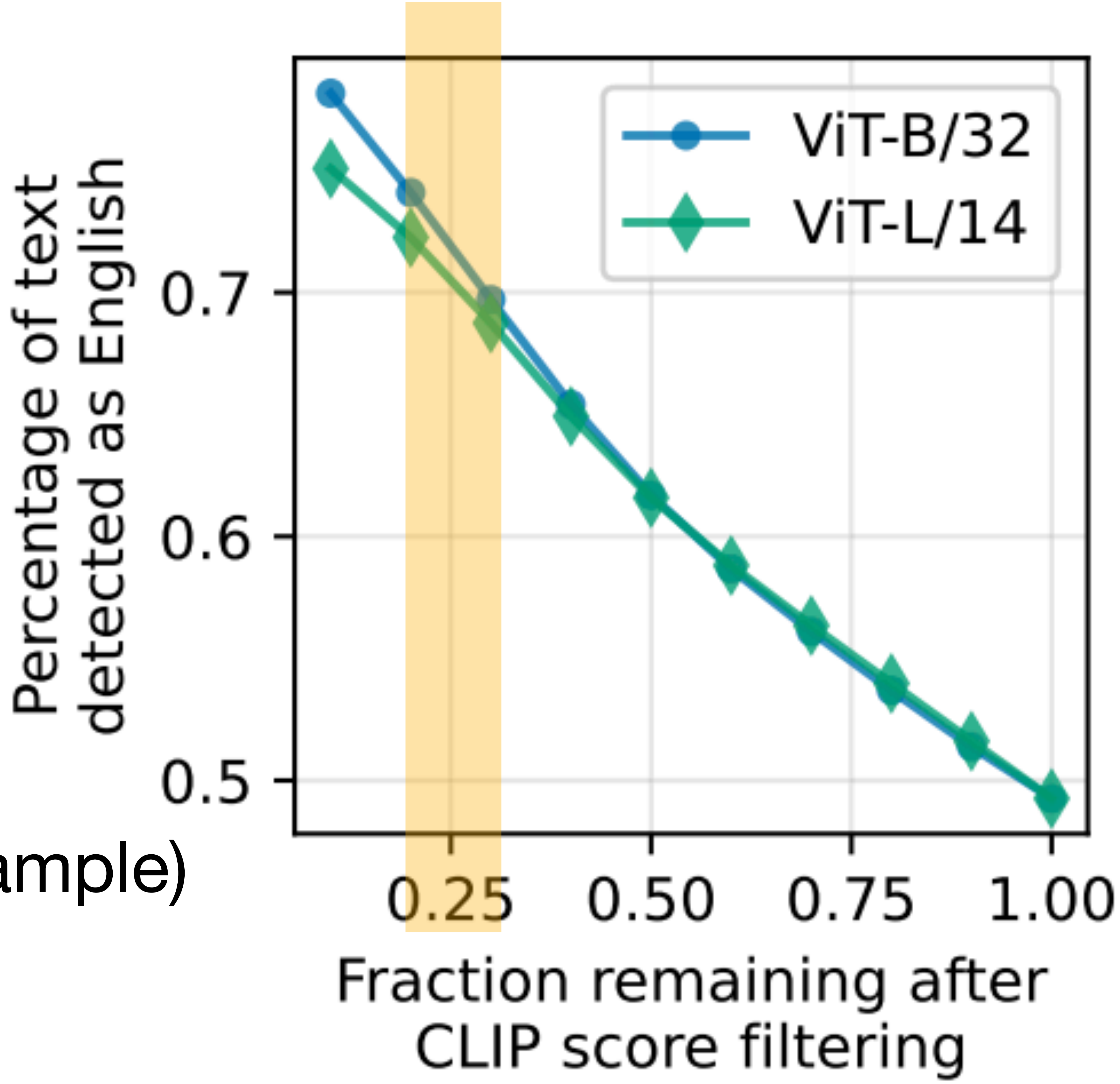
Gadre et al. “Datacomp: In search of the next generation of multimodal datasets”, NeurIPS 2023

Monolingual bias in DataComp pipeline



Monolingual bias in DataComp pipeline

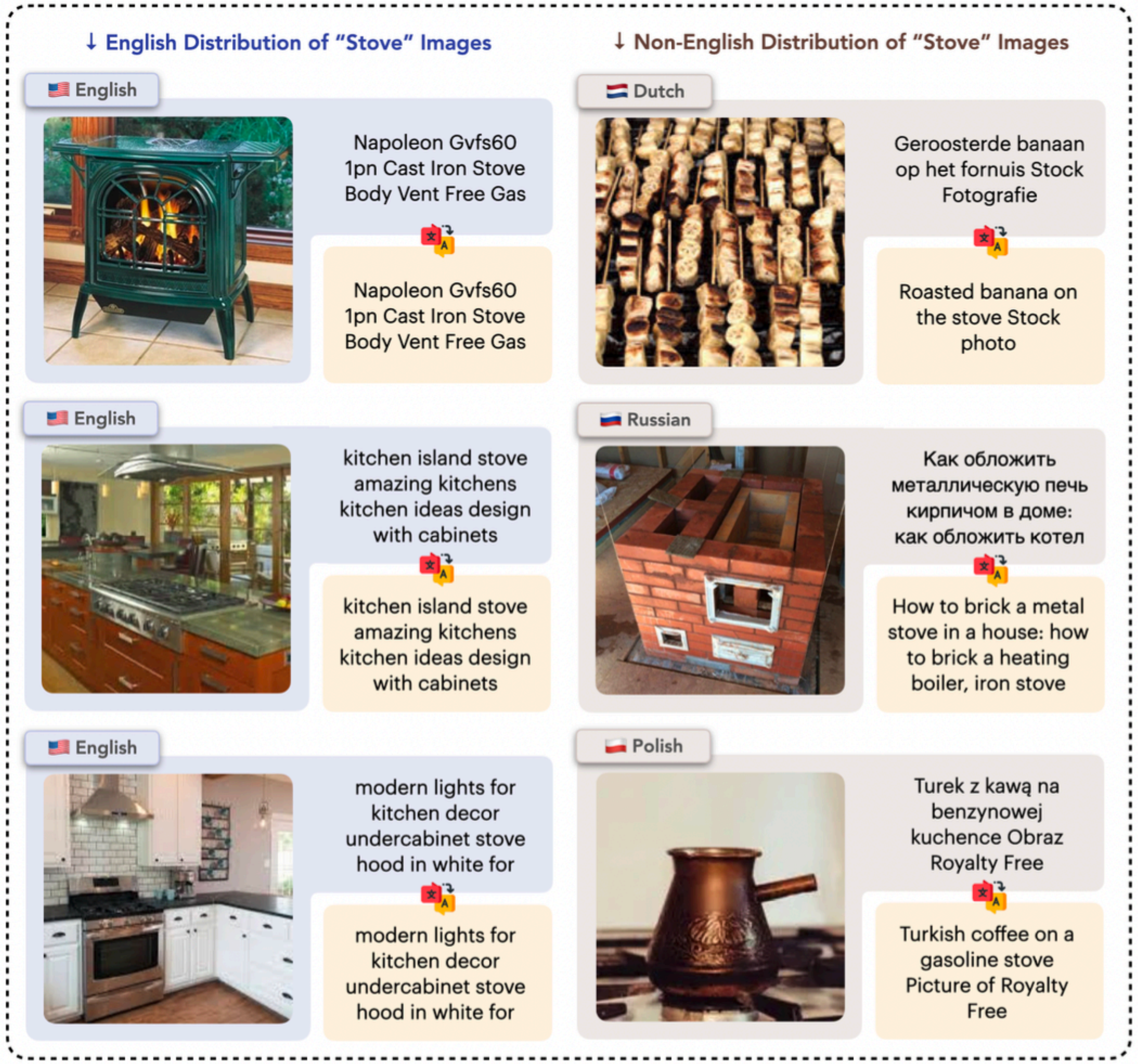
- Selection bias in DataComp, LAION, etc.
- Why do the filters have strong English bias?
  - because of dataset bias, human bias, and evaluation benchmark bias
- Challenges
  - **image** diversity is lost (stove example)
  - **text** diversity is lost (Japanese caption example)

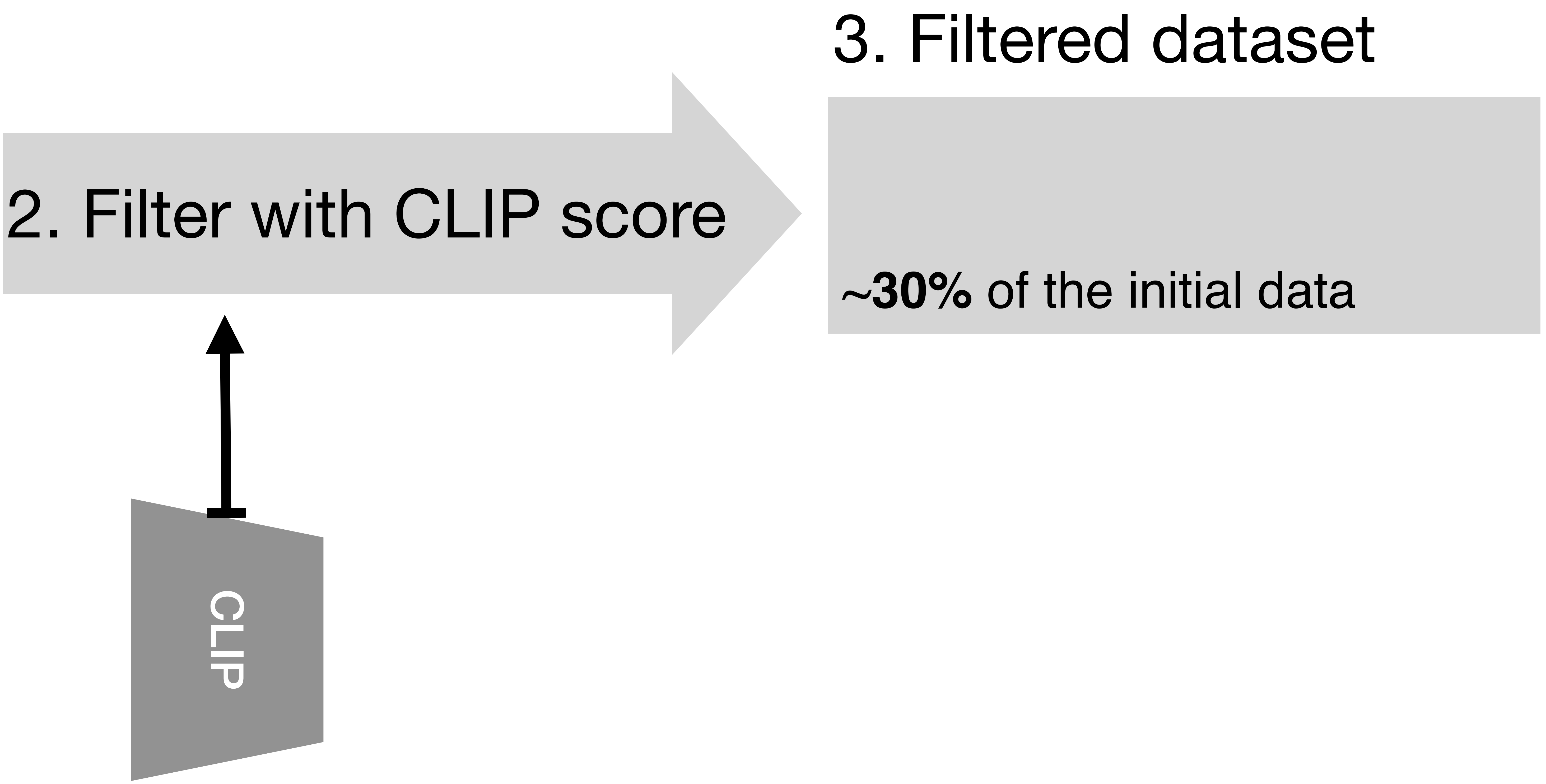


Fang, et al. “Data filtering networks”, ICLR 2024

Can we fight the monolingual bias baked into the data curation pipeline?

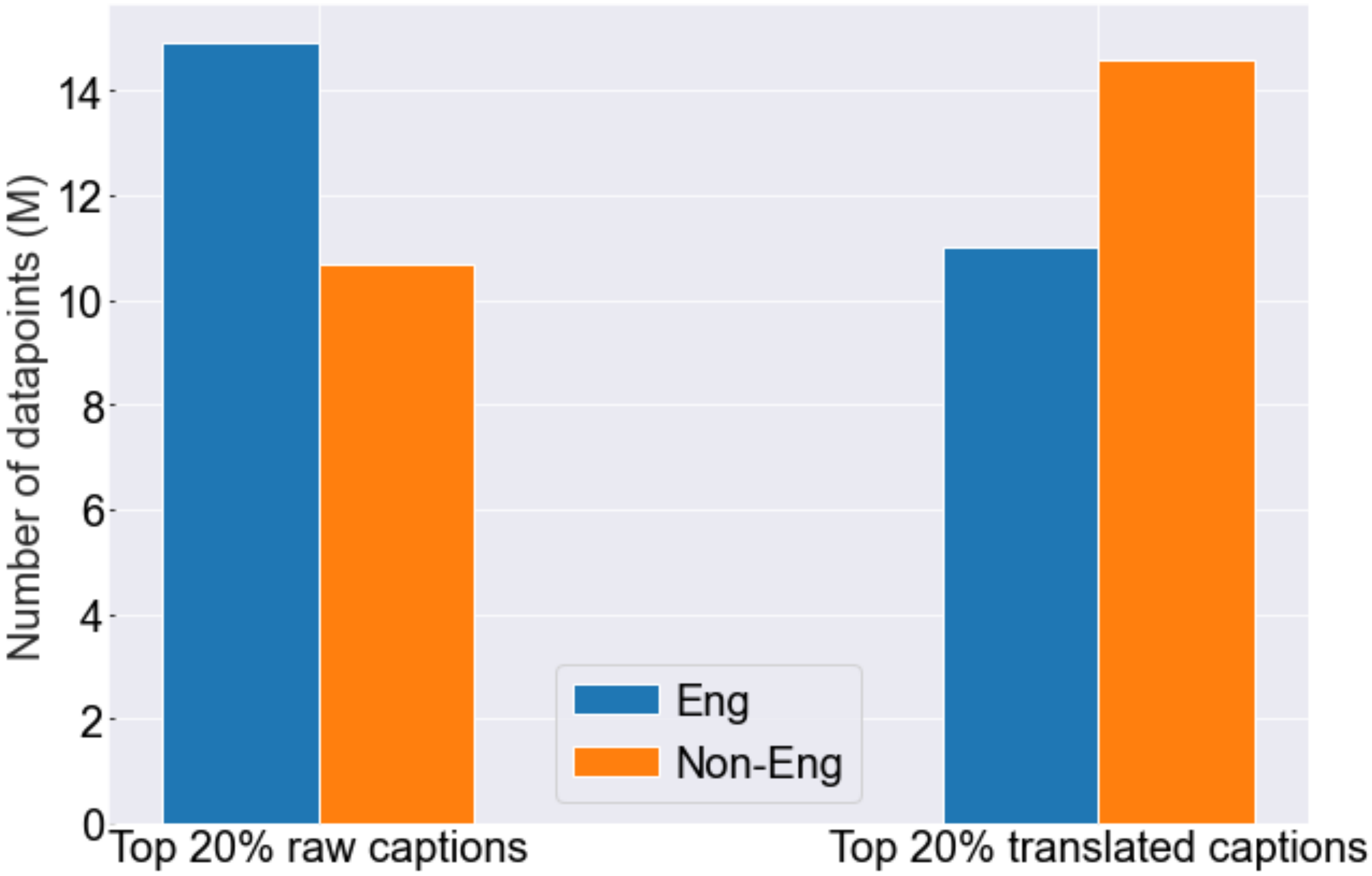
- Hypothesis:
  Separate the semantic diversity in images, and semantic diversity in text from the language it is written in, using translation models:





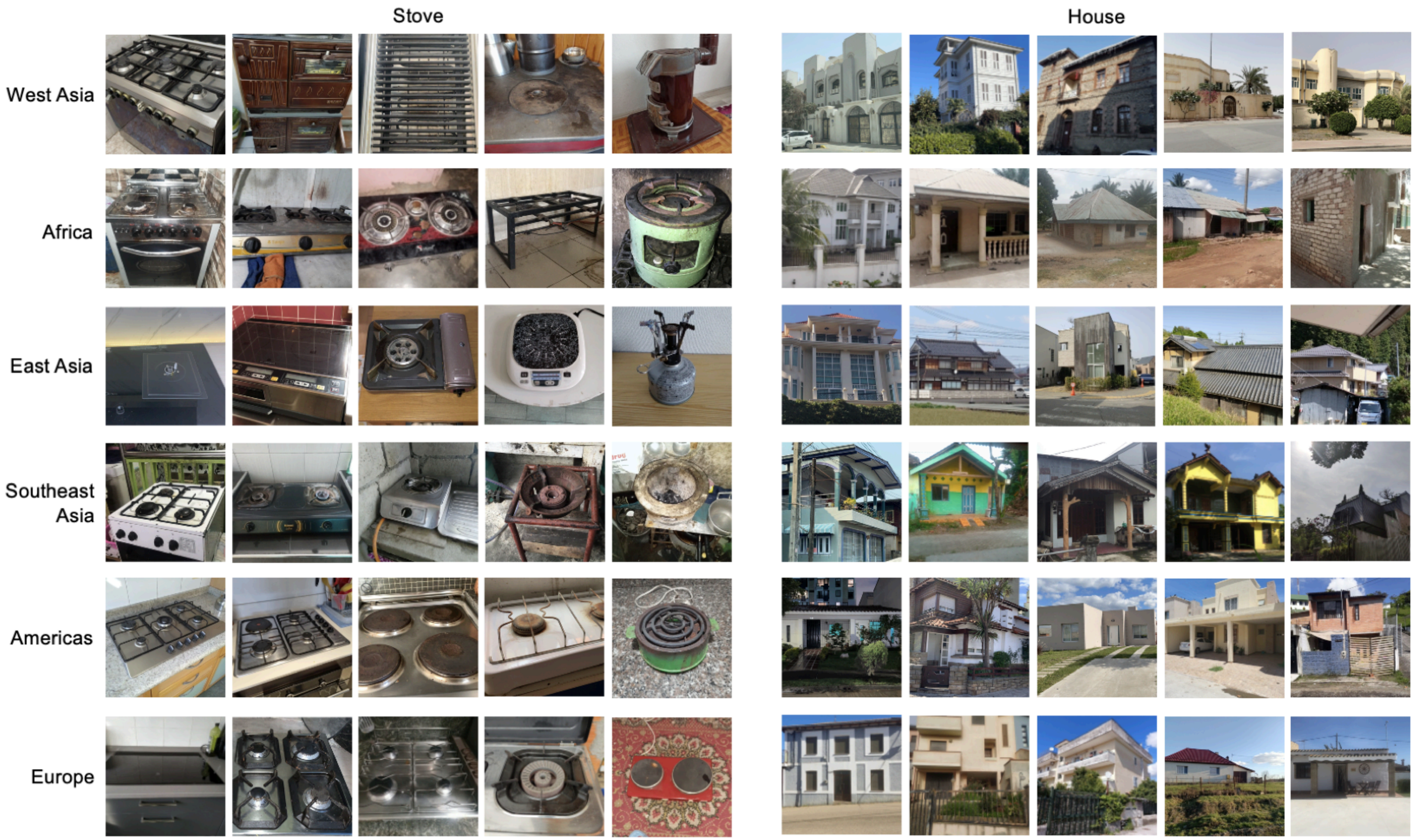
Translating captions to English increases the CLIP score

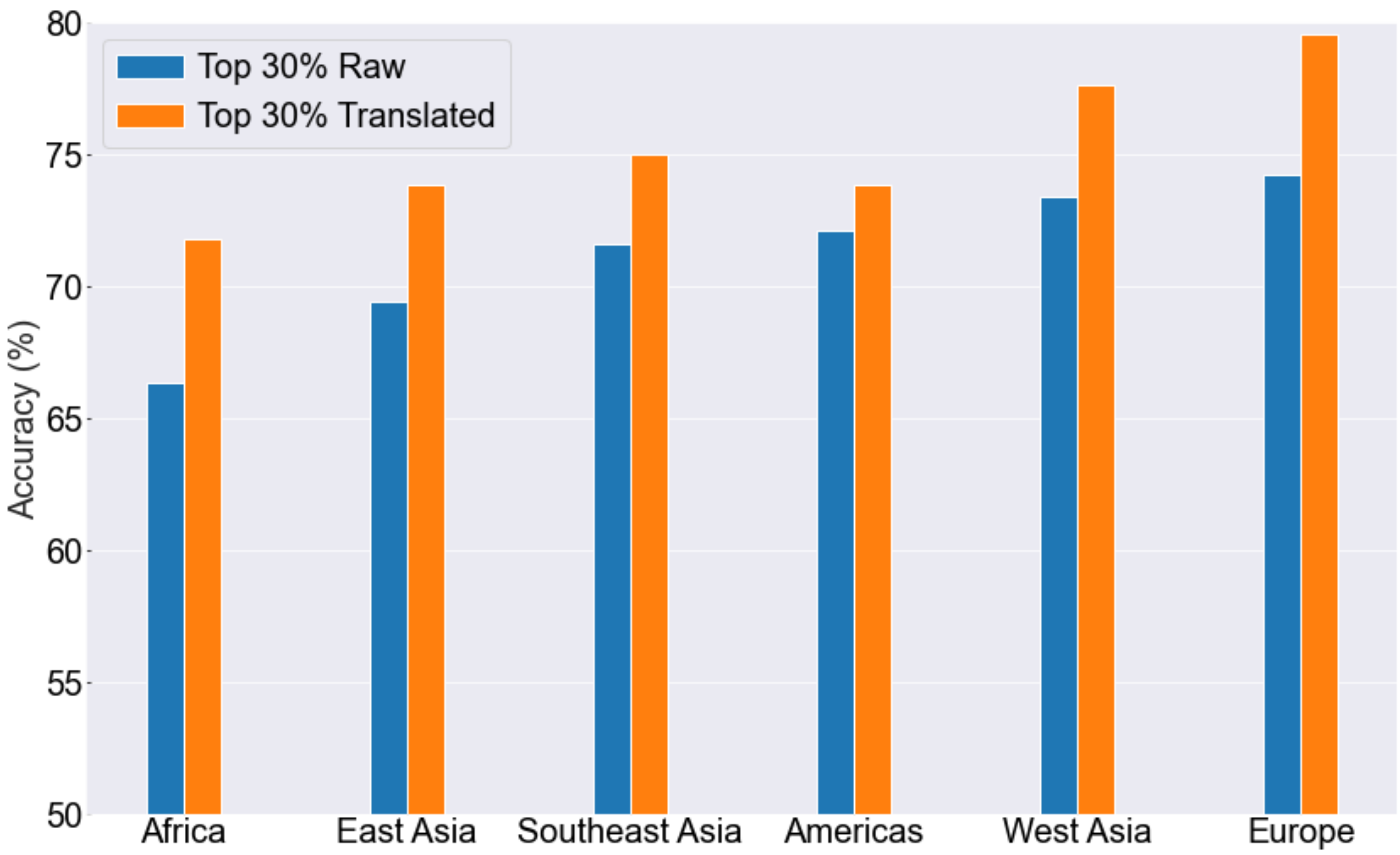
- When 25.6M=20% of data pass the filter



Evaluation benchmarks designed for inclusiveness

- **GeoDE**: classes are standard but images are diverse

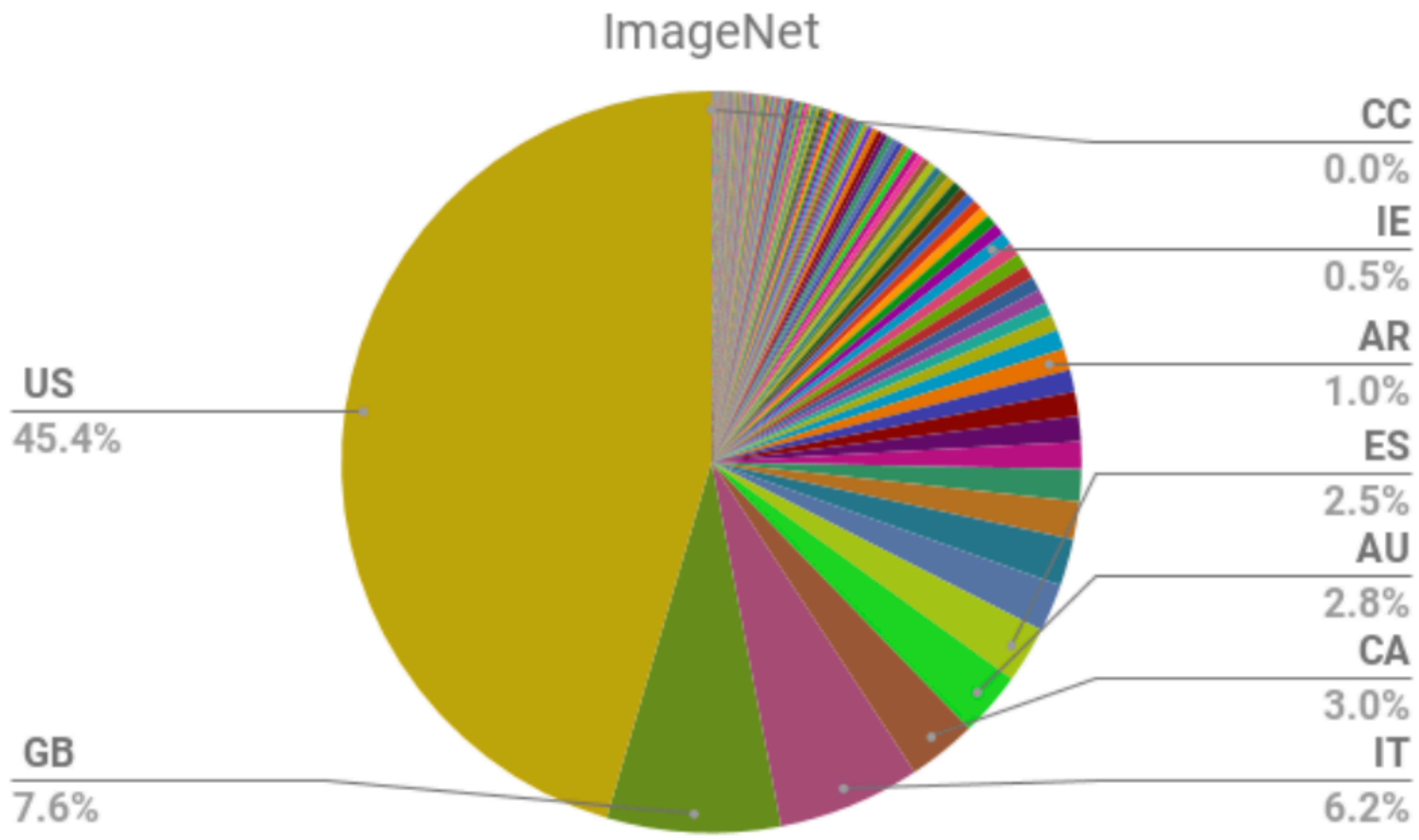




- The diversity in images help on inclusive benchmarks

Ramaswamy et al. "GeoDE: a Geographically Diverse Evaluation Dataset for Object Recognition", NeurIPS 2023

ImageNet is known to have English bias



- Still, perhaps surprisingly, using diverse training data performs best

Baselines (1.28B samples seen)	ImageNet	IN Shifts	Retrieval	38-task Average
Filtered raw captions	0.414	0.340	0.344	0.414
Filtered translated captions	0.427	0.347	0.352	0.414
Filtered raw captions + Filtered translated captions	**0.456**	**0.369**	**0.371**	**0.435**

Is the gain coming from diverse images or captions?

- Ideally, we would like to fix a dataset of diverse images, and have
  - dataset A captioned by both English speaking and Non-English speaking crowd, and
  - dataset B captioned by English speaking crowd

	Baseline name	Dataset size	ImageNet	ImageNet shifts	GeoDE	Average over 38 tasks
dataset A	Top 20% translated captions	25.6M	0.329	0.275	0.709	0.359
dataset B	Top 20% translated captions, replaced with synthetic captions	25.6M	0.283	0.255	0.696	0.336

Li, Li, Savarese, Hoi. “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.”, ICML 2023

Takeaways

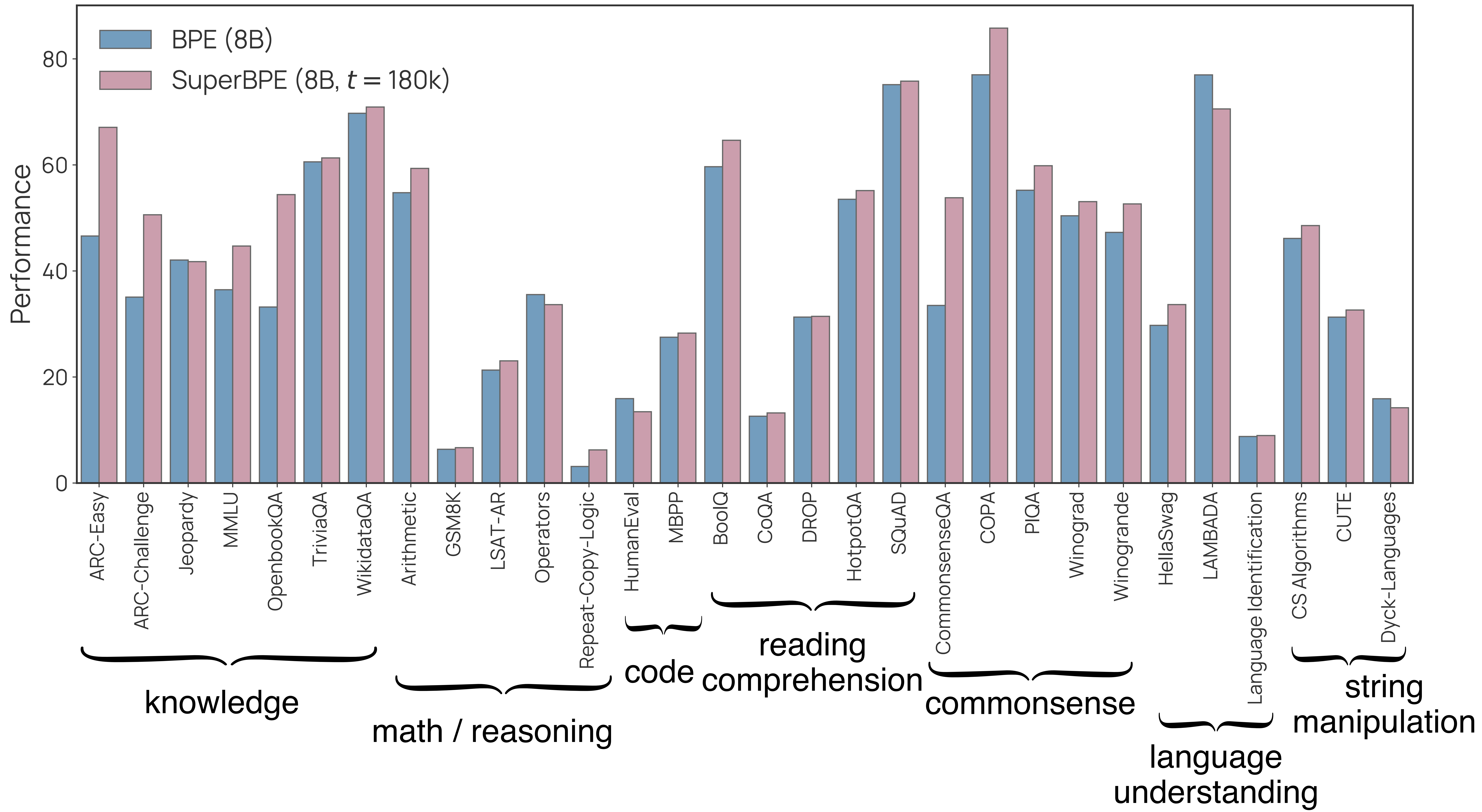
- DataComp and LAION pipeline suffers from English bias
- By translating internet-scale image-text data before filtering, we improve both the image diversity and text diversity, which gives overall improved performance on downstream tasks
- More experiments are needed to generalize our findings to other filtering algorithms, other translation methods, and other non-English downstream tasks

References

- “**SuperBPE: Space Travel for Language Models**”, Alisa Liu, Jonathan Hayase, Valentin Hofmann, Sewoong Oh, Noah A. Smith, Yejin Choi, https://arxiv.org/pdf/2503.13423,

- “**Multilingual diversity improves vision-language representations**”, Thao Nguyen, Matthew Wallingford, Sebastin Santy, Wei-Chiu Ma, Sewoong Oh, Ludwig Schmidt, Pang Wei W Koh, Ranjay Krishn, *NeurIPS 2024*
    - “Datacomp: In search of the next generation of multimodal datasets”, NeurIPS 2023
    - “Improving Multimodal Datasets with Image Captioning”, NeurIPS 2023

SuperBPE downstream performance



Baseline of translating already filtered dataset

Baseline name	Dataset size	ImageNet	ImageNet shifts	Retrieval	GeoDE	Average over 38 tasks
Training with DataComp setup (128M steps)						
Filtered raw captions	25.6M	0.316	0.260	0.282	0.688	0.350
Filtered raw captions, replaced with translated captions	25.6M	0.304	0.252	0.268	0.668	0.331

M. Nezhurina, R. Beaumont, R. Vencu, and C. Schuhmann. Laion translated: 3b captions translated to english from laion5b.