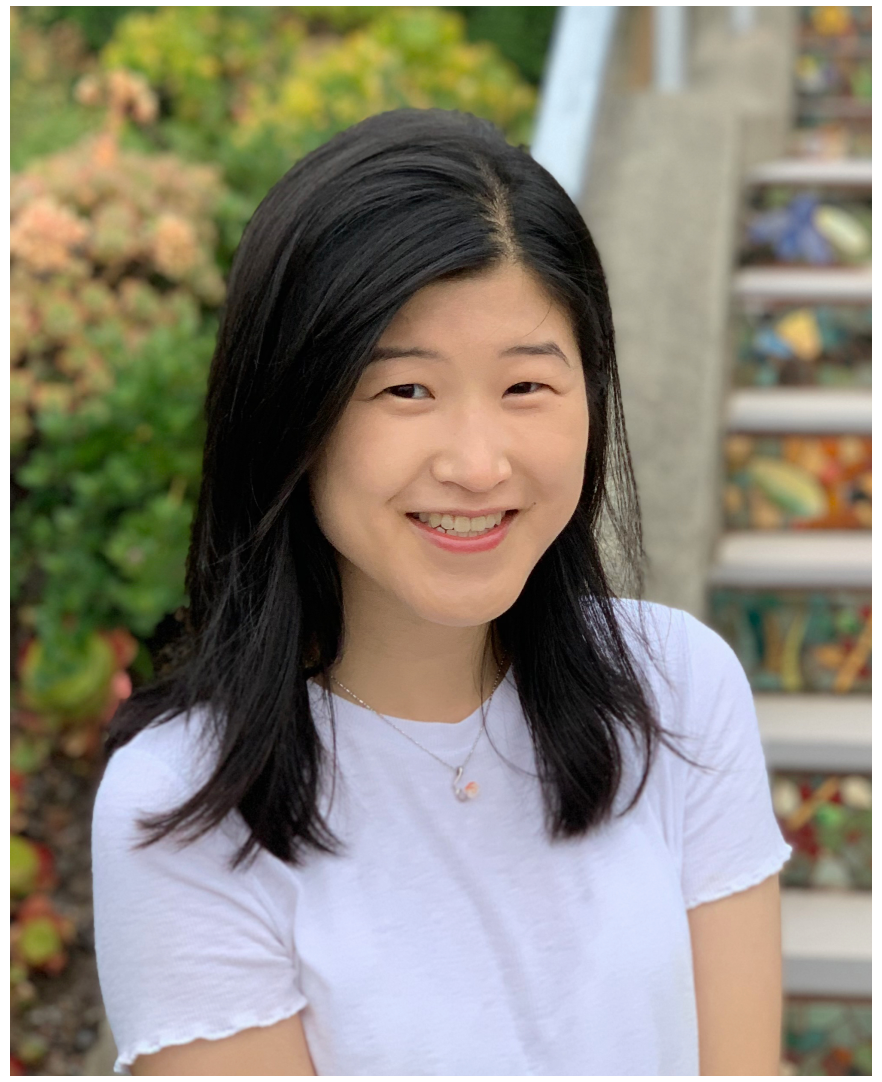
Superword tokenizer for LLMs & dataset mixture inference

AIMACCS @ OSU 2025

Sewoong Oh, University of Washington

* Alisa Liu



* Jonathan Hayase

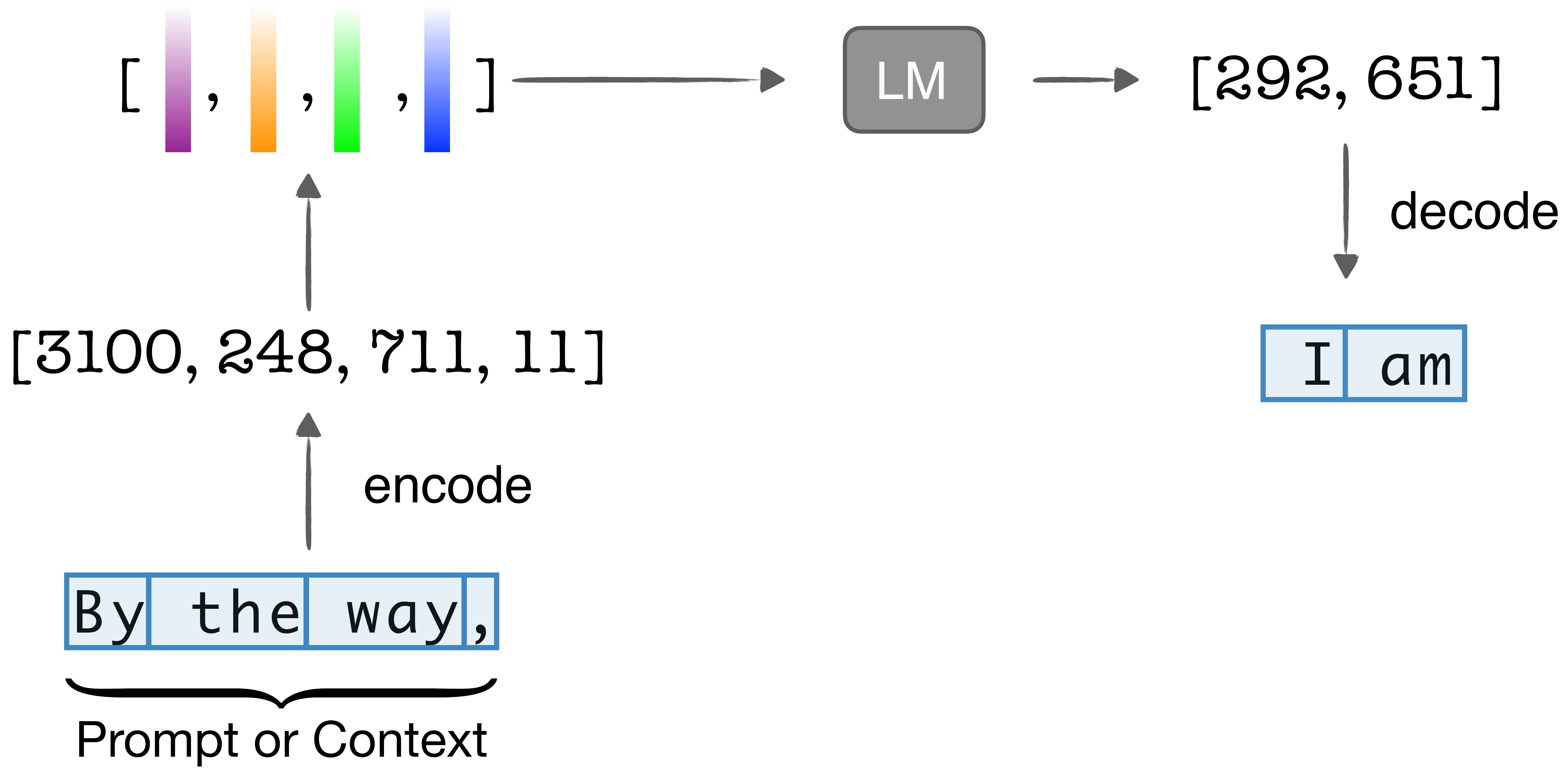
Valentin Hofmann

Noah Smith

Yejin Choi

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:

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	.

  - **Efficiency is bad**: the number of tokens needed to represent text is quite large, which increases the input dimension of the model

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.

  - Much more efficient:
    - about 5 characters in English word on average
  - but can encounter new words that is not in the vocab, which is represented by a special token **<UNK>,** since there are many more uncommon words

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

_ 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

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

$\boxed{\text{_ t}}$	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

$\boxed{\text{_ t}}$	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	Pair counts		Vocabulary
_ P r o o f , _ o f , _t he ,	i n	8024060	_t
_ M i l k y , _ W a y , _ c o			_a
n s i s t i n g , _ o f , _ m			he
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	e r	6517430	
i l e i , _ u s e d , _a , _t	r e	6031043	
e l e s c o p e , _t o , _ s	o n	5261131	
t u d y , _t he , _ M i l k	_ i	5209828	
y , _ W a y , _a n d , _ d i s	_ o	5163783	
c o v e r e d , _t h a t , _	_ s	5035505	
i t , _ i s , _ c o m p o s e	_ w	4523998	
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 .			

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

$\boxed{\text{i n}}$	$\boxed{8024060}$
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

$\boxed{\text{i n}}$	$\boxed{8024060}$
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

$\boxed{\text{\_t he}}$	$\boxed{5605612}$
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

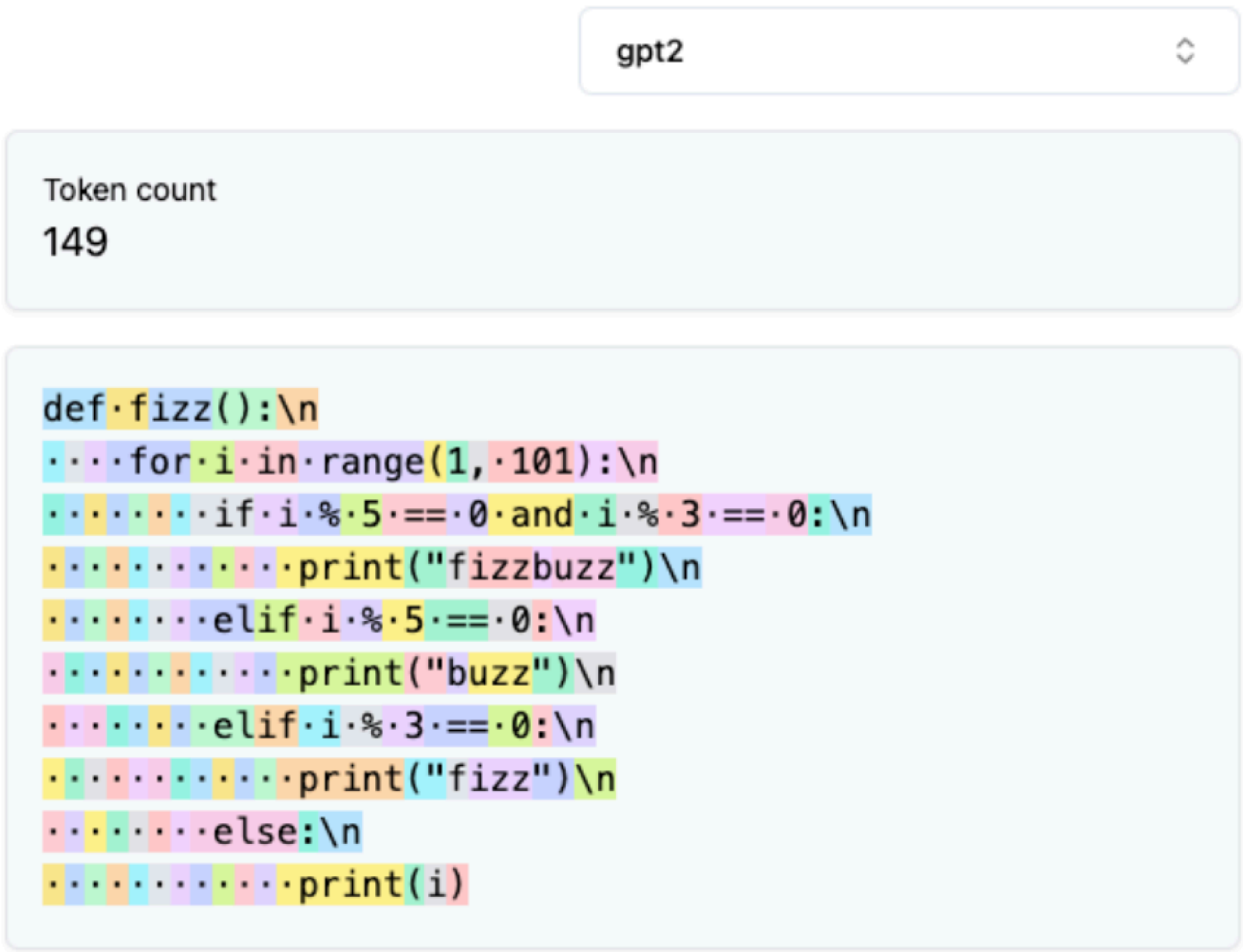
_the

⋮

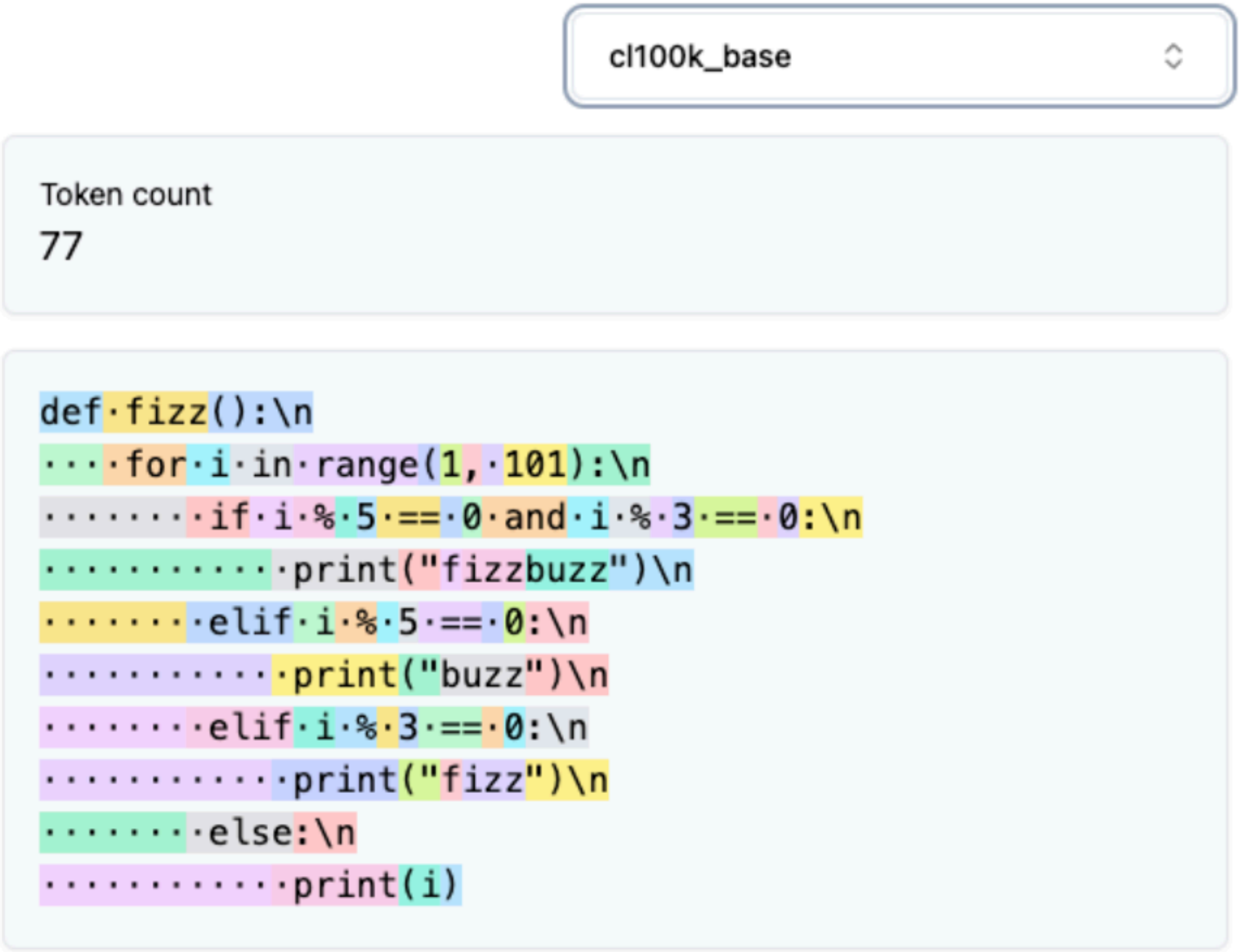
until we reach desired vocab size T

Trade-off between vocab size and efficiency

GPT-2 Tokenizer: vocab size 50k and not trained on coding data



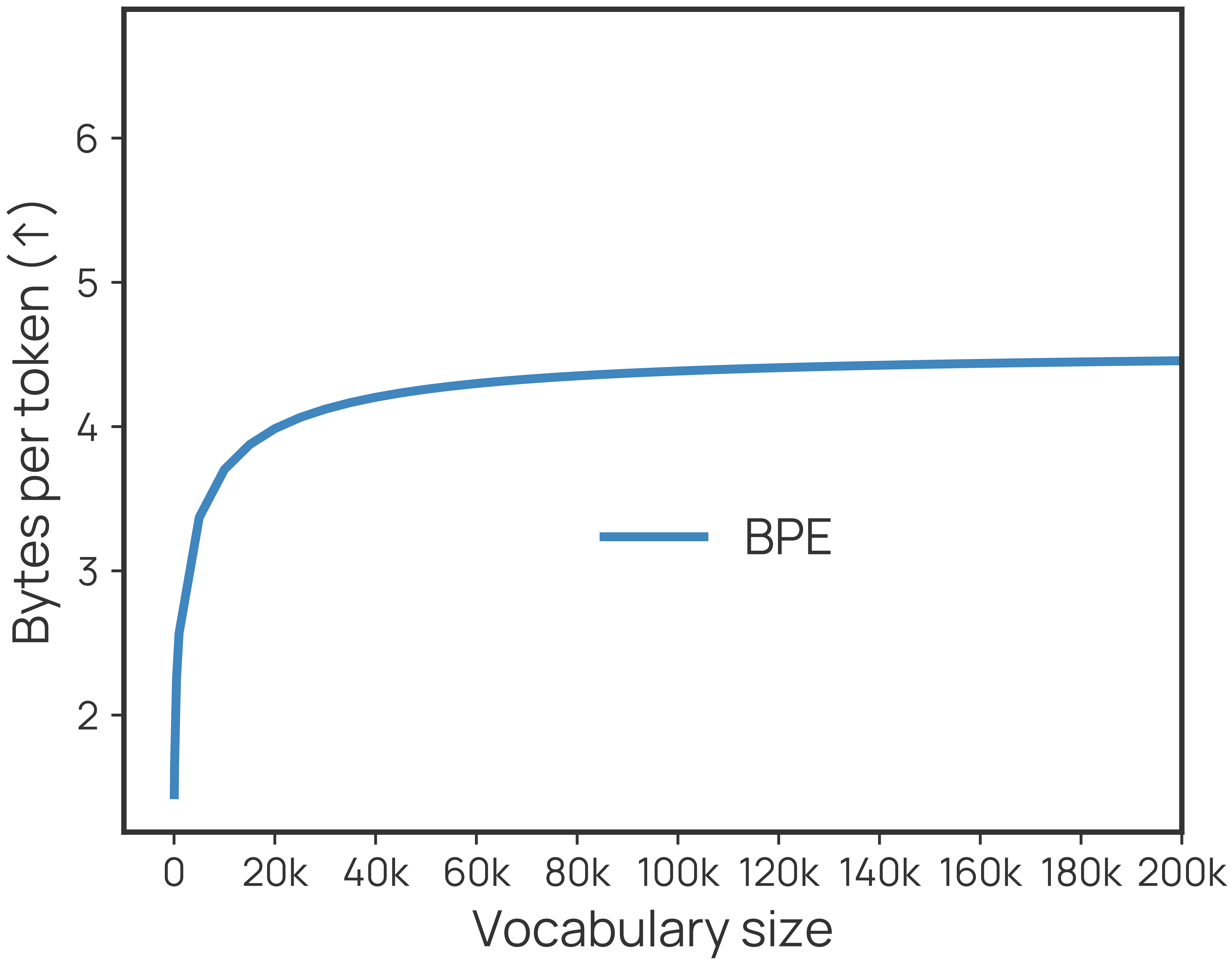
GPT-4 Tokenizer: vocab size 100k and trained on coding data



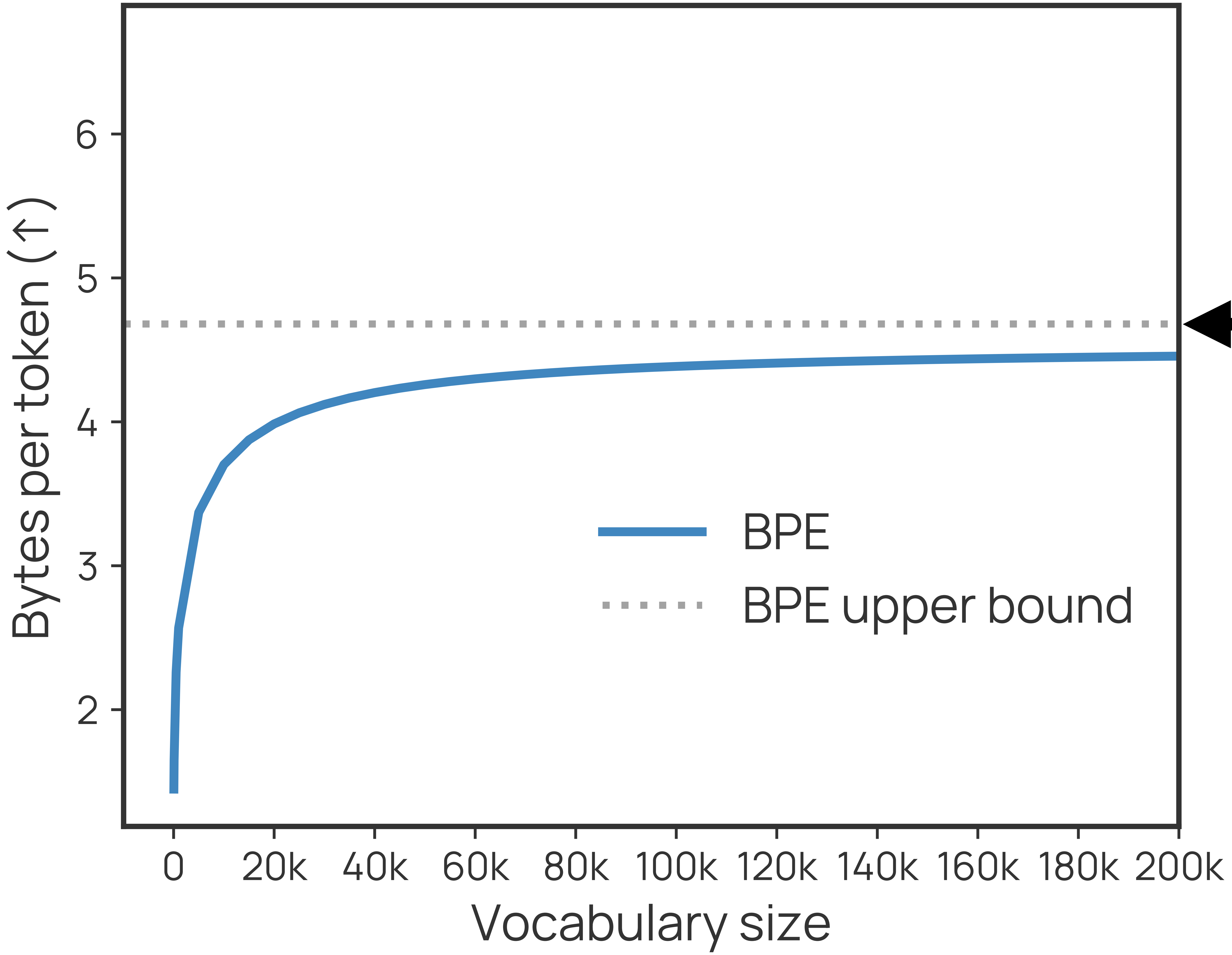
- Why can we not arbitrarily increase the vocab size?
- Question: How do we know what training data these closed-source tokenizers are trained on? [“Data Mixture Inference Attack: BPE Tokenizers Reveal Training Data Compositions”, *NeurIPS 2024*]

courtesy of https://github.com/openai/tiktoken

Trade-off between vocab size and efficiency



Fundamental limit of subword tokenization



Average number of characters in an English word

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

$\boxed{\text{\_of \_the}}$	$\boxed{517482}$
' 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

$\text{stage 1} \left\{ \begin{array}{l} \text{\_t} \\ \text{\_a} \\ \text{he} \\ \text{in} \\ \text{re} \\ \text{\_the} \\ \vdots \\ \text{\_Aleg} \end{array} \right.$

_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

$\boxed{\text{_in _the}}$	$\boxed{362529}$
' 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

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

Vocabulary

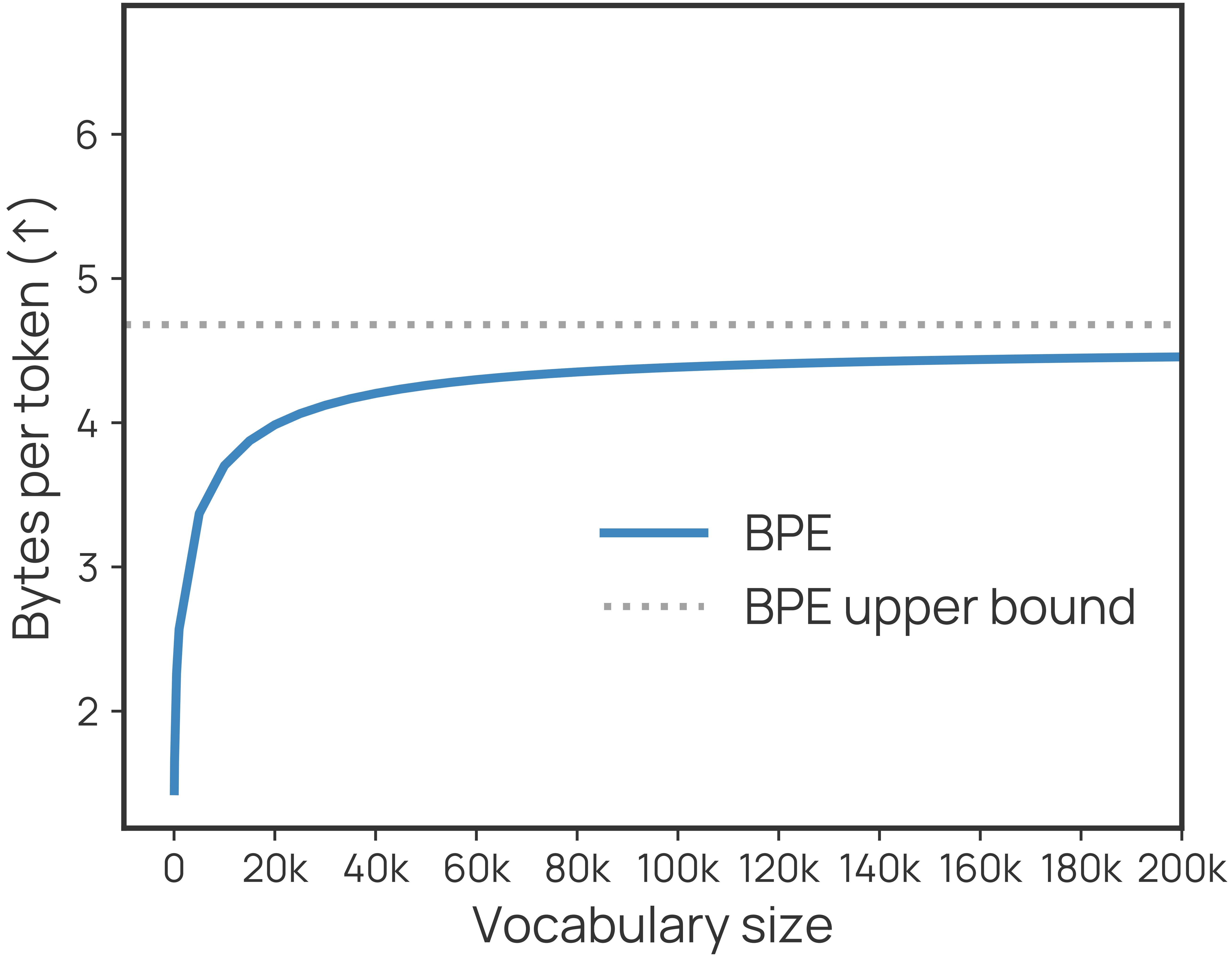
stage 1

_t
_a
he
in
re
_the
⋮
_Aleg

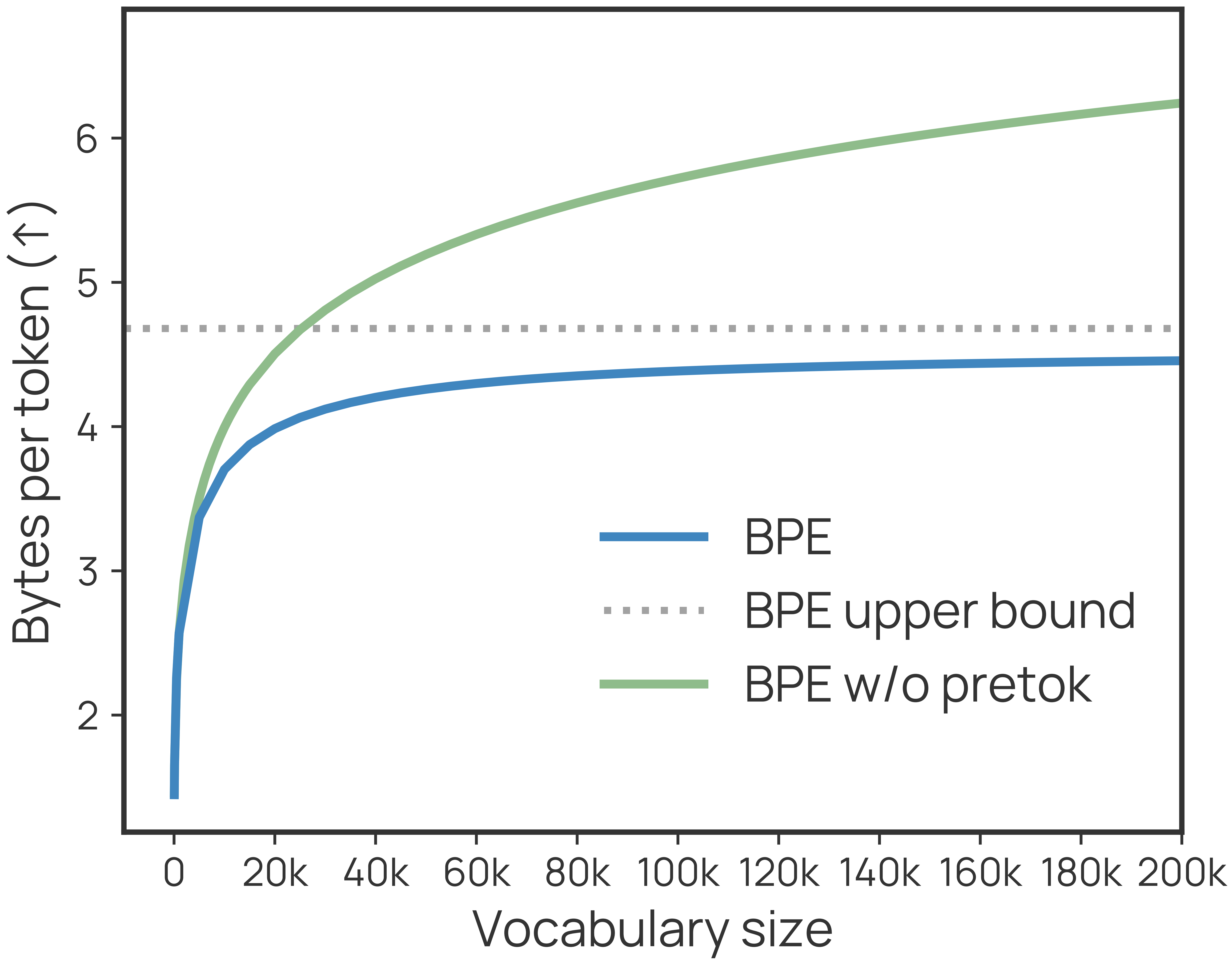
_of _the
' s
, _and
_in _the
⋮

until we reach desired vocab size T

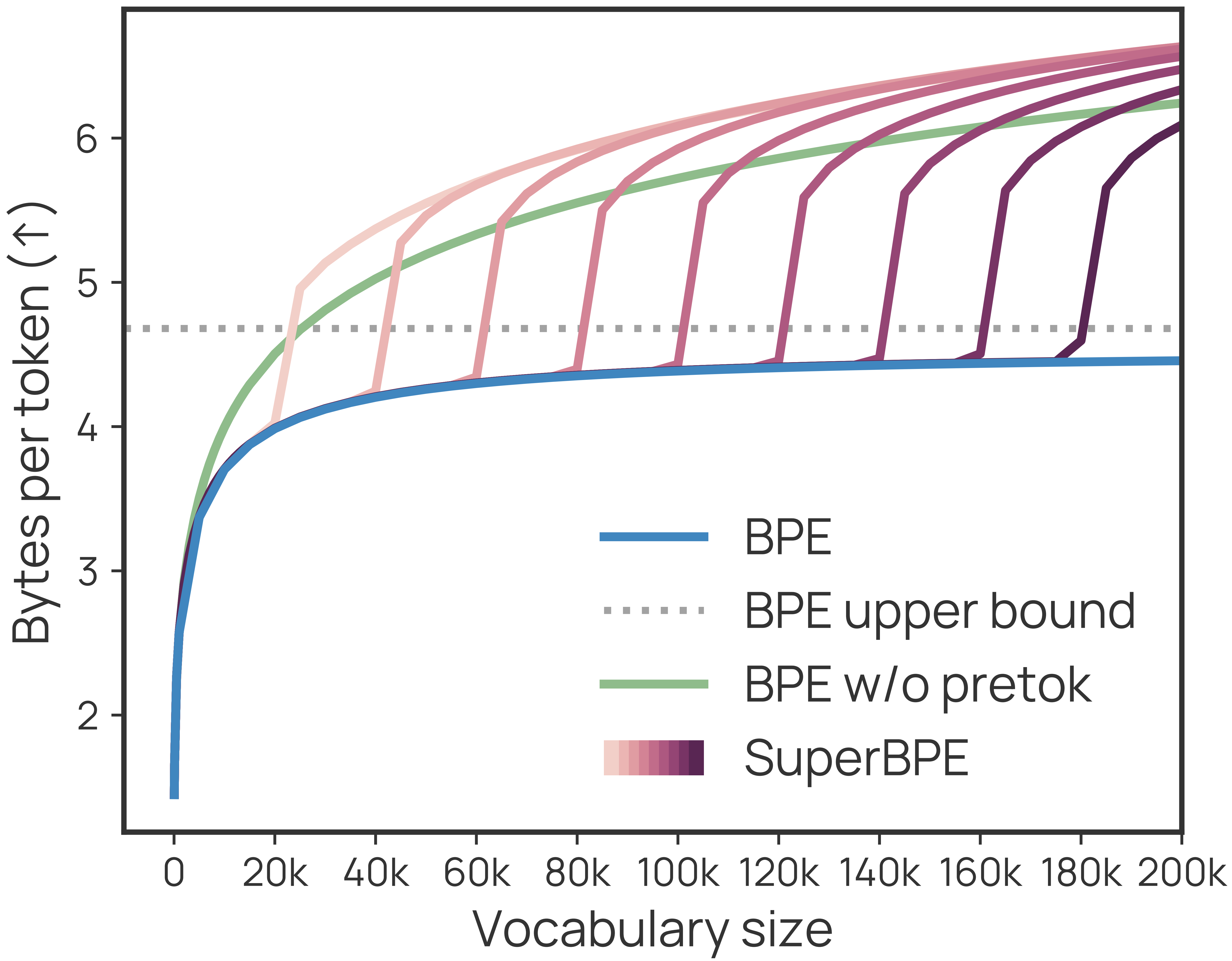
SuperBPE encodes text more efficiently



SuperBPE encodes text more efficiently



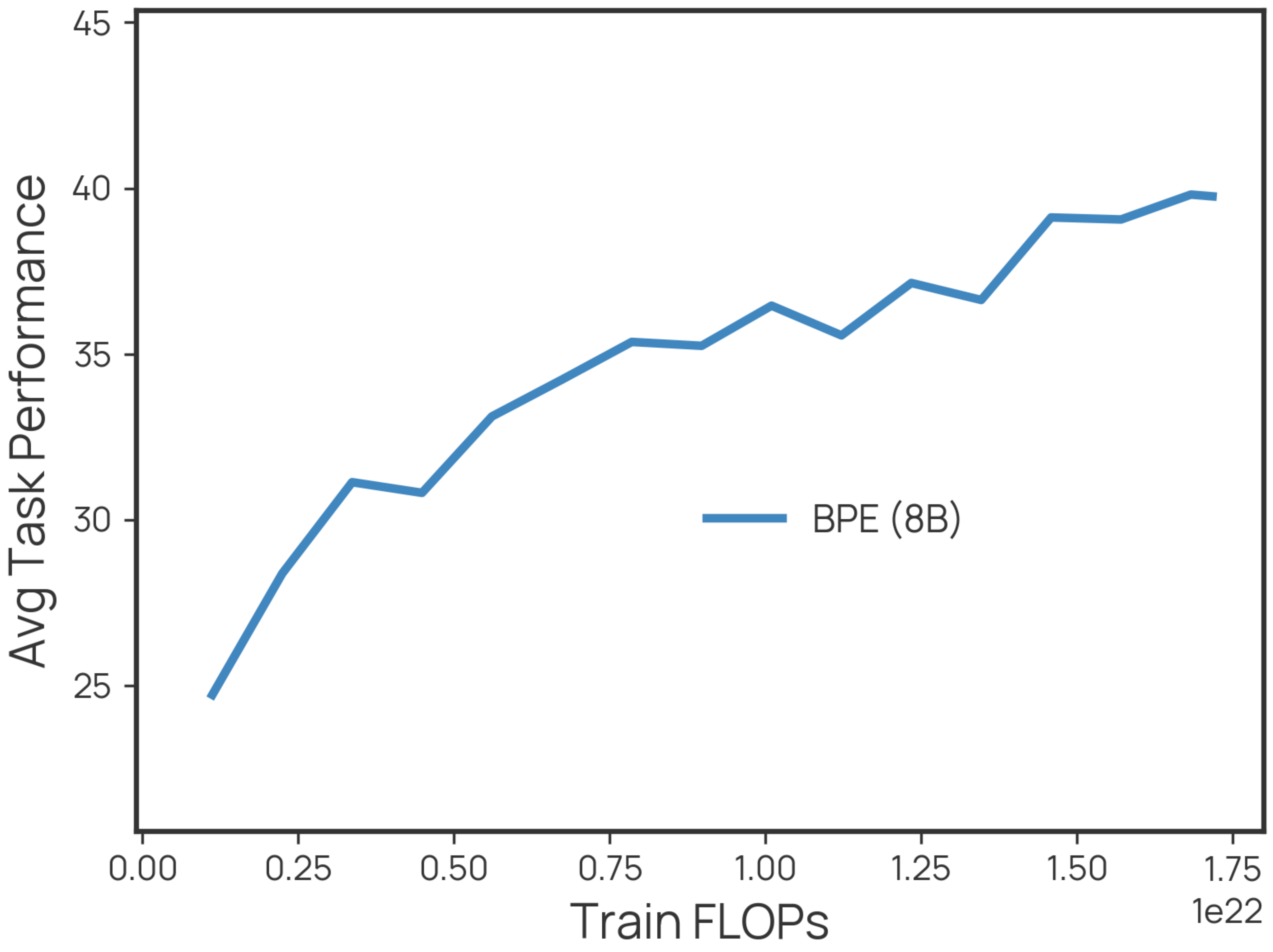
SuperBPE encodes text more efficiently



Changing tokenizer requires pretraining LLM

Baseline:

- Tokenizer: **BPE** with 200k tokens
- Model size: **8B** parameters
- Train data: **330B** tokens from OLMO2
- Evaluation
  - Average performance on 30 tasks

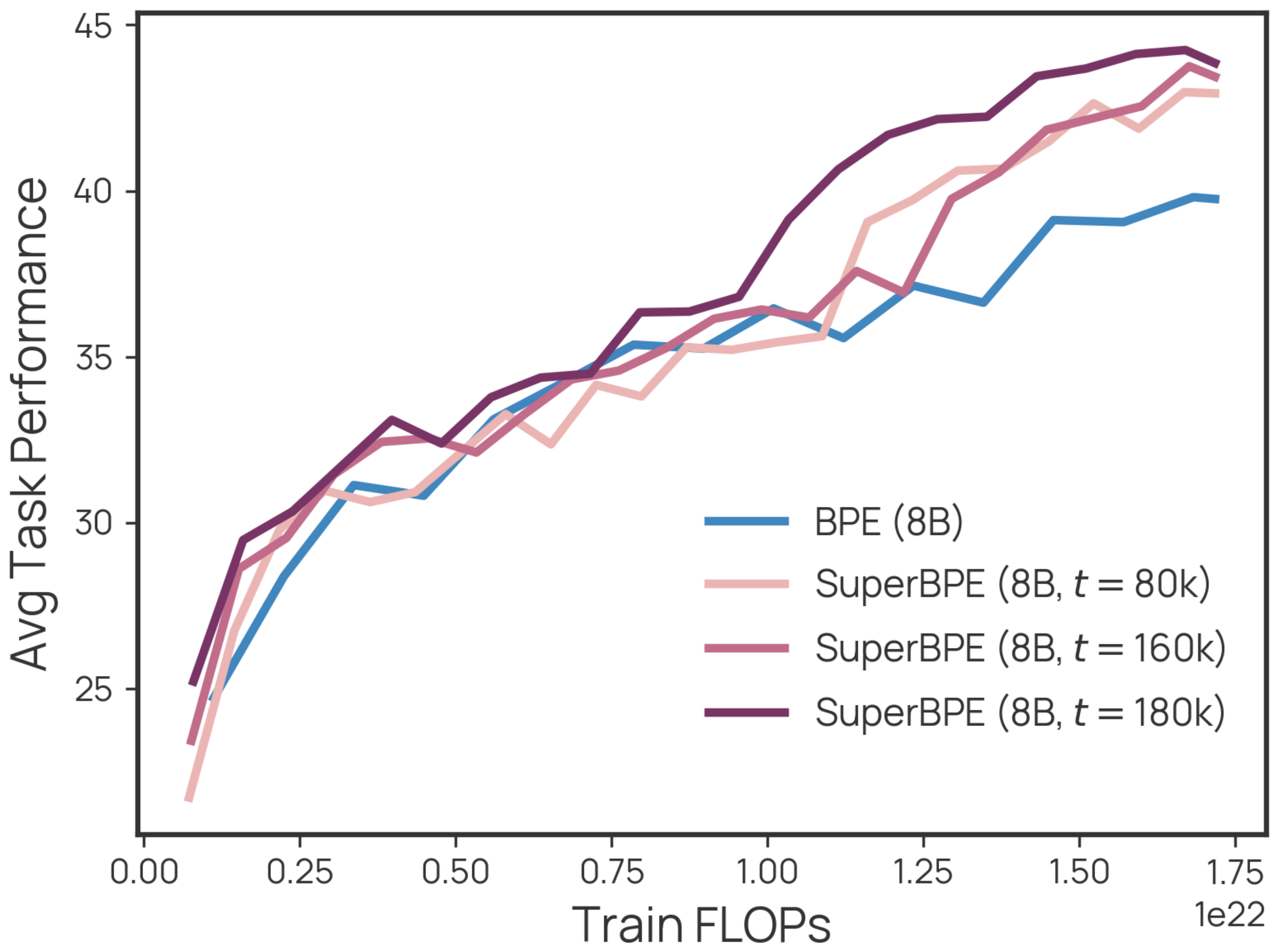


In a fair comparison, SuperBPE outperforms in 30 downstream tasks

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

SuperBPE 8B

- ✅ model size **8B**
- ✅ train data **330B tokens** OLMO2
- ✅ training compute is the same
- ❌ inference compute (**35% less**)
- ❌ amount of text seen (**41% more**)



Is this fair?

	model size	*	train tokens	=	train compute
BPE	**8B**		330B		1.75e22 FLOPs
SuperBPE	**8B**		330B		1.75e22 FLOPs

	train tokens	*	Bytes per Token	=	train text
BPE	330B		4.5		1485B
SuperBPE	330B		6.1		2013B

Is this fair?

	model size *	train tokens =	train compute
BPE	8B	330B	1.75e22 FLOPs
SuperBPE	11B	243B	1.75e22 FLOPs

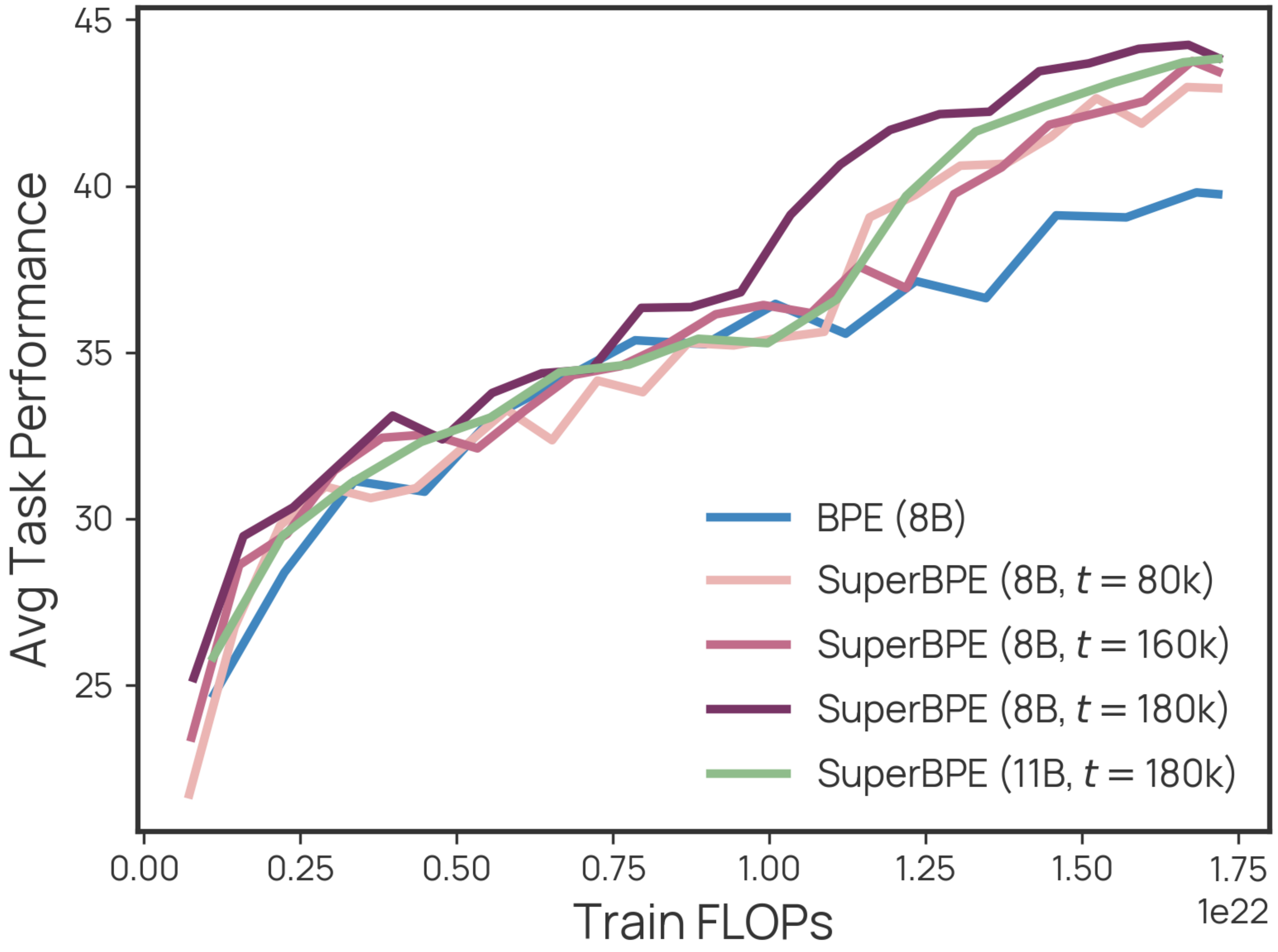
	train tokens *	Bytes per Token =	train text
BPE	330B	4.5	1485B
SuperBPE	243B	6.1	1485B

In a fair comparison, SuperBPE outperforms in 30 downstream tasks

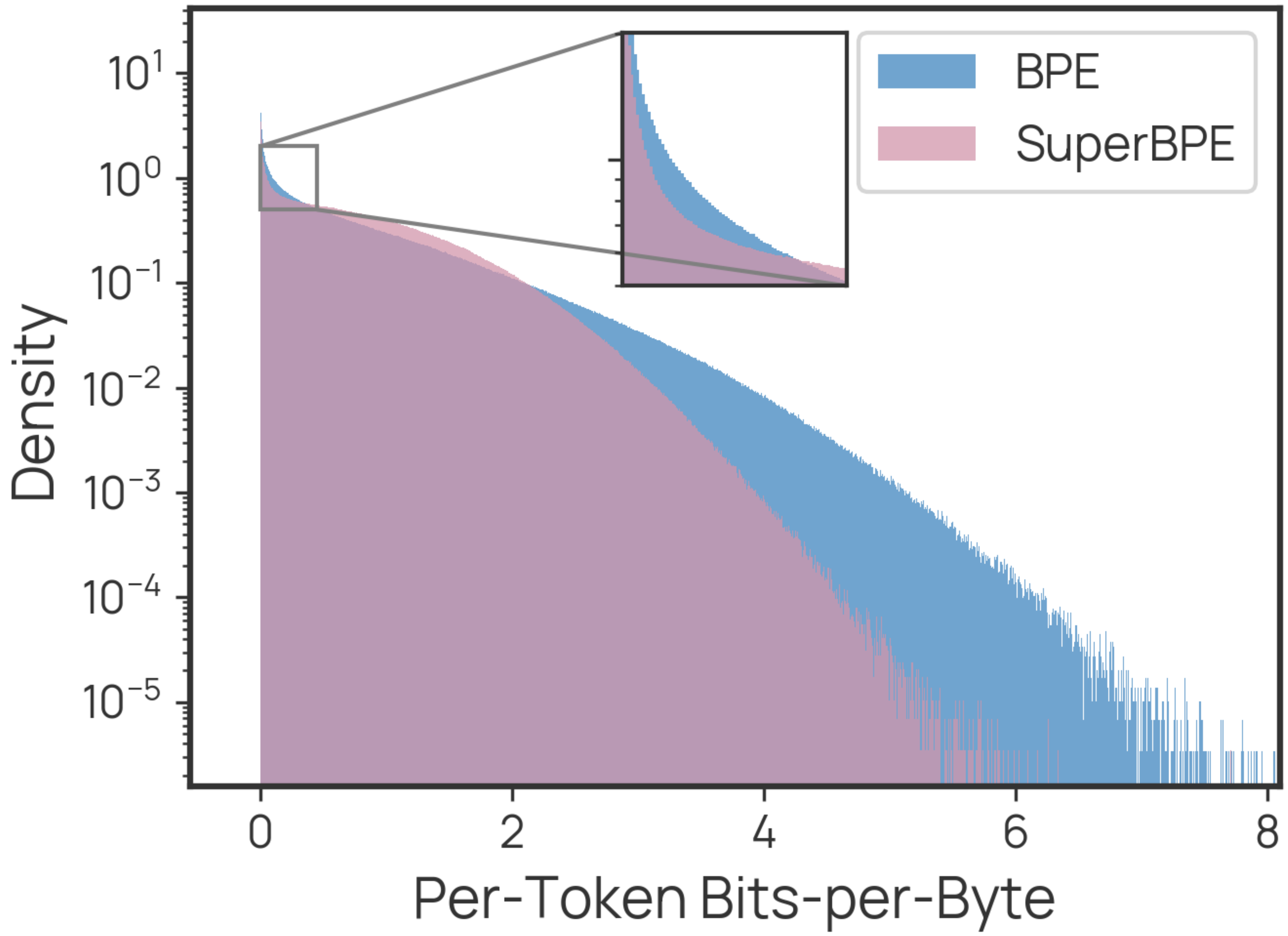
Baseline: **BPE 8B** (Olmo2 @ 330B tokens)

SuperBPE 11B

- ❌ model size **11B** (39% bigger)
- ✅ training data **330B** tokens
- ✅ train compute is the same
- ✅ inference compute: same
- ✅ amount of text seen: same



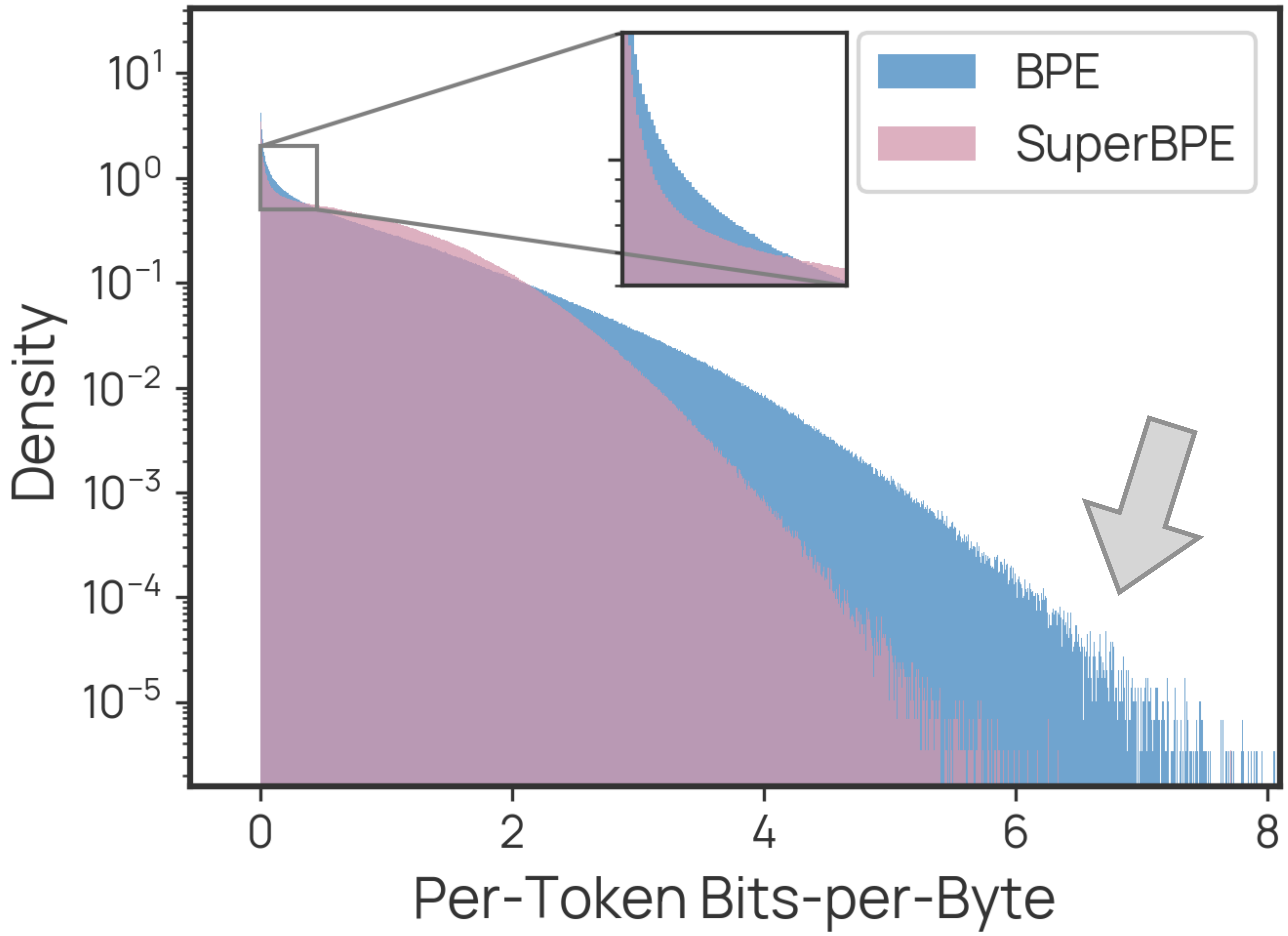
BPB Distribution



Mean BPB is very close (SuperBPE behind by 0.0017)

But SuperBPE distributes loss *more uniformly* over tokens

BPB Distribution



SuperBPE has fewer very-high-loss tokens:

May explain why we win on evals (evals are hard)

Even after models plateau in loss, they keep getting better at evals

Same Pretraining Loss, Better Downstream (Liu et al., 2023)

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

In this era of data-centric AI, pretaining data for LLMs is a trade secret

- Typical attacks to reveal something about the pertaining (or fine-tuning) data attempt to identify the membership, i.e., answer a question like
  “Is Harry Potter used in training?”
  due to its importance in privacy and copyright.

- This is a very challenging question with mixed results:

Prompt (taken from article):
WASHINGTON — American intelligence officials have concluded that a Russian military intelligence unit secretly offered bounties to Taliban-linked militants for killing coalition

Output from GPT-4:
forces in Afghanistan — including targeting American troops — amid the peace talks to end the long-running war there, according to officials briefed on the matter.
The United States concluded months ago that the Russian unit, which has been linked to assassination attempts and other covert operations in Europe intended to destabilize the West or take revenge on turncoats, had covertly offered rewards for successful attacks last year.
Islamist militants, or armed criminal elements closely associated with them, are believed to have collected some bounty money, the officials said. Twenty Americans were killed in combat in Afghanistan in 2019, but it was not clear which killings were under suspicion.
The intelligence finding was briefed to President Trump, and the White House’s National Security Council discussed the problem at an interagency meeting in late March, the officials said. Officials developed a menu of potential options — starting with making a diplomatic complaint to Moscow and a demand it stop, along with an escalating series of sanctions and other possible responses, but the White House has yet to authorize any step, the officials

Actual text from NYTimes:
forces in Afghanistan — including targeting American troops — amid the peace talks to end the long-running war there, according to officials briefed on the matter.
The United States concluded months ago that the Russian unit, which has been linked to assassination attempts and other covert operations in Europe intended to destabilize the West or take revenge on turncoats, had covertly offered rewards for successful attacks last year.
Islamist militants, or armed criminal elements closely associated with them, are believed to have collected some bounty money, the officials said. Twenty Americans were killed in combat in Afghanistan in 2019, but it was not clear which killings were under suspicion.
The intelligence finding was briefed to President Trump, and the White House’s National Security Council discussed the problem at an interagency meeting in late March, the officials said. Officials developed a menu of potential options — starting with making a diplomatic complaint to Moscow and a demand that it stop, along with an escalating series of sanctions and other possible responses, but the White House has yet to authorize any step, the officials

Membership inference
from Language Models

**Dataset mixture inference
from BPE tokenizers**

Data Mixture Inference

English $\mathcal{D}_{\text{En}}$

Normalize the digits, then ensure that they sum to 1.

Python $\mathcal{D}_{\text{Py}}$

```
x = logits.softmax()  # get probs
assert x.sum().item() == 1  # compare
```

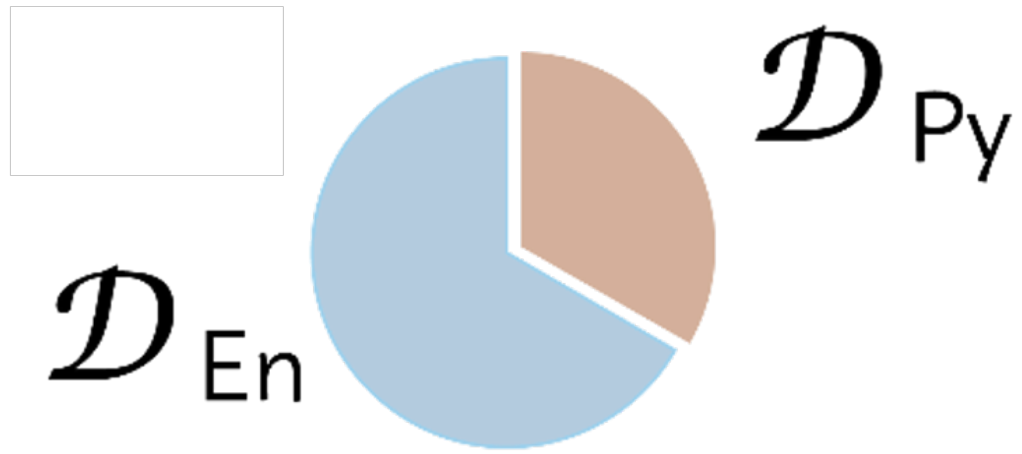
Data Mixture Inference

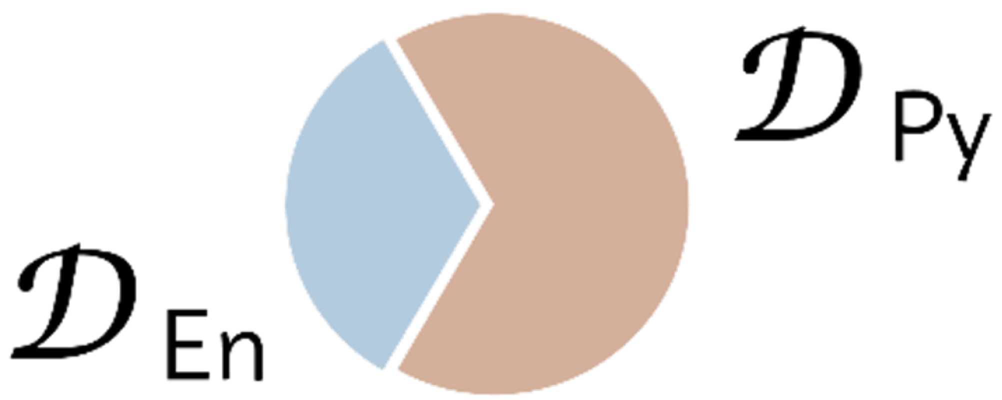
English $\mathcal{D}_{\text{En}}$

Normalize **the** dig**it**s, **the**n ensure **th**at **th**ey sum to 1.

Python $\mathcal{D}_{\text{Py}}$

```
x = logits.softmax()   # get probs
assert x.sum().item() == 1  # compare
```





Given data, BPE learns a merge list

1	_ t
2	_t h
3	_th e
4	s u
5	i t
6	( )

1	( )
2	i t
3	_ t
4	s u
5	_ #
6	_ =

merge list

The learned merge list is (very) sensitive to the mixture ratio of data distributions

Data Mixture Inference

English $\mathcal{D}_{\text{En}}$

Normalize the digits, then ensure that they sum to 1.

Python $\mathcal{D}_{\text{Py}}$

```
x = logits.softmax()  # get probs
assert x.sum().item() == 1  # compare
```

merge list

1	_ t
2	_t h
3	_th e
4	s u
5	i t
6	( )

1	( )
2	i t
3	_ t
4	s u
5	_ #
6	_ =

Given a merge list, can we solve for the mixture ratio?

The learned merge list is (very) sensitive to the mixture ratio of data distributions

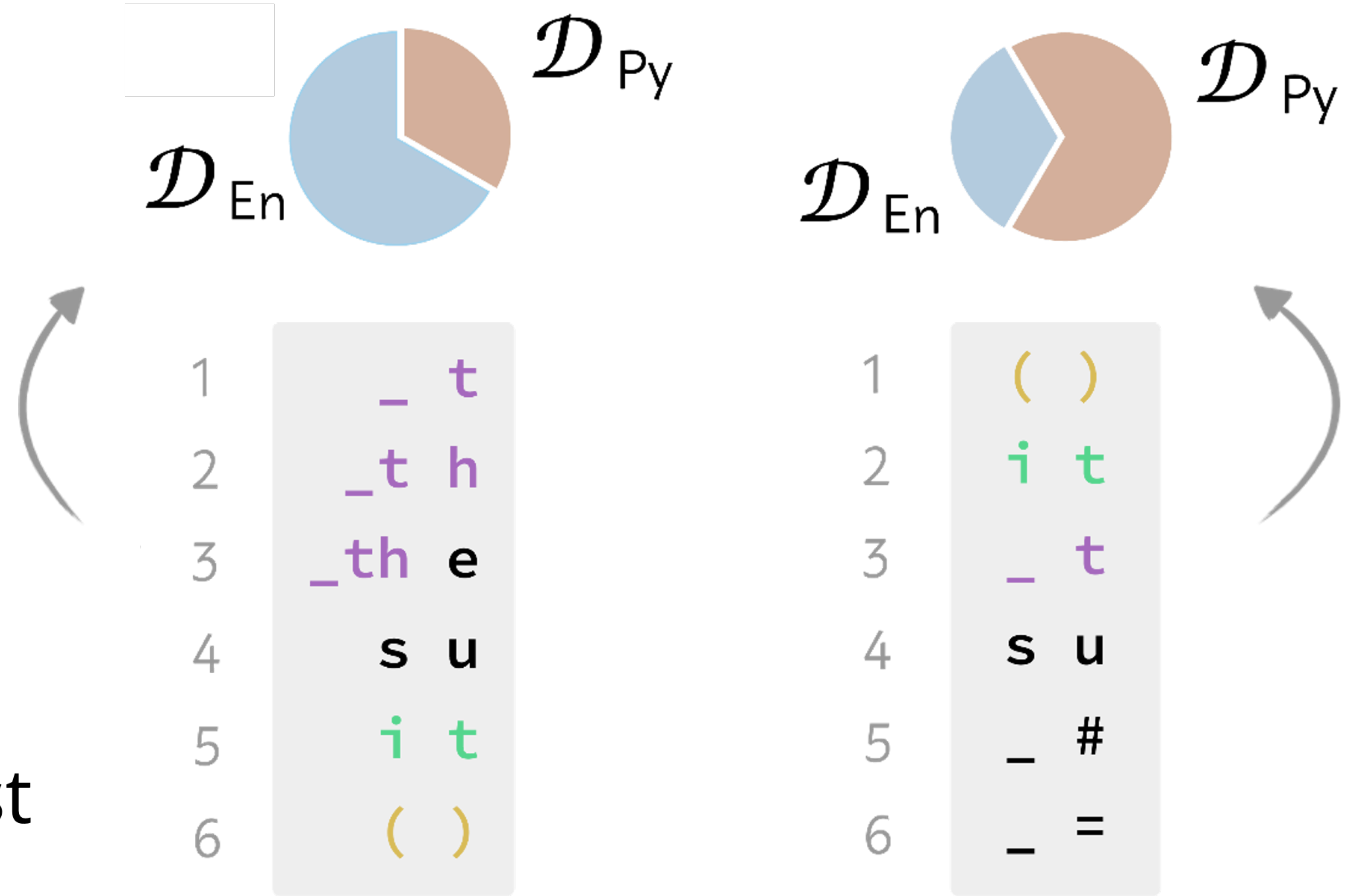
Data Mixture Inference

English $\mathcal{D}_{\text{En}}$

Normalize the digits, then ensure that they sum to 1.

Python $\mathcal{D}_{\text{Py}}$

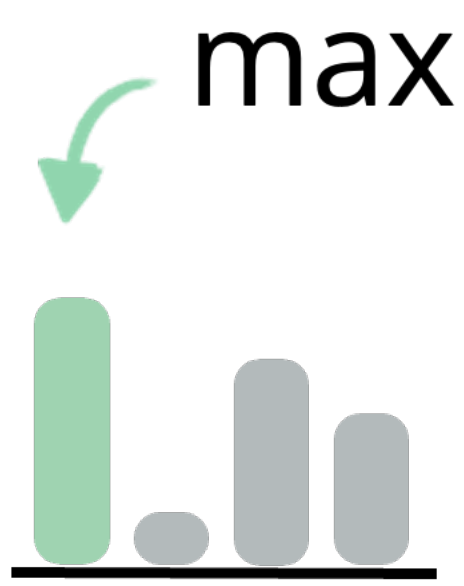
```
x = logits.softmax()  # get probs
assert x.sum().item() == 1  # compare
```



merge list

Given a merge list, can we solve for the mixture ratio?

Merge List — counts in En corpus — counts in Py corpus

> 1 _ _ $\alpha^{En}\,[\;6 \quad 1 \quad 1774 \quad 2246\;] + \alpha^{Py}\,[\;13055 \quad 248 \quad 1083 \quad 648\;] =$ 

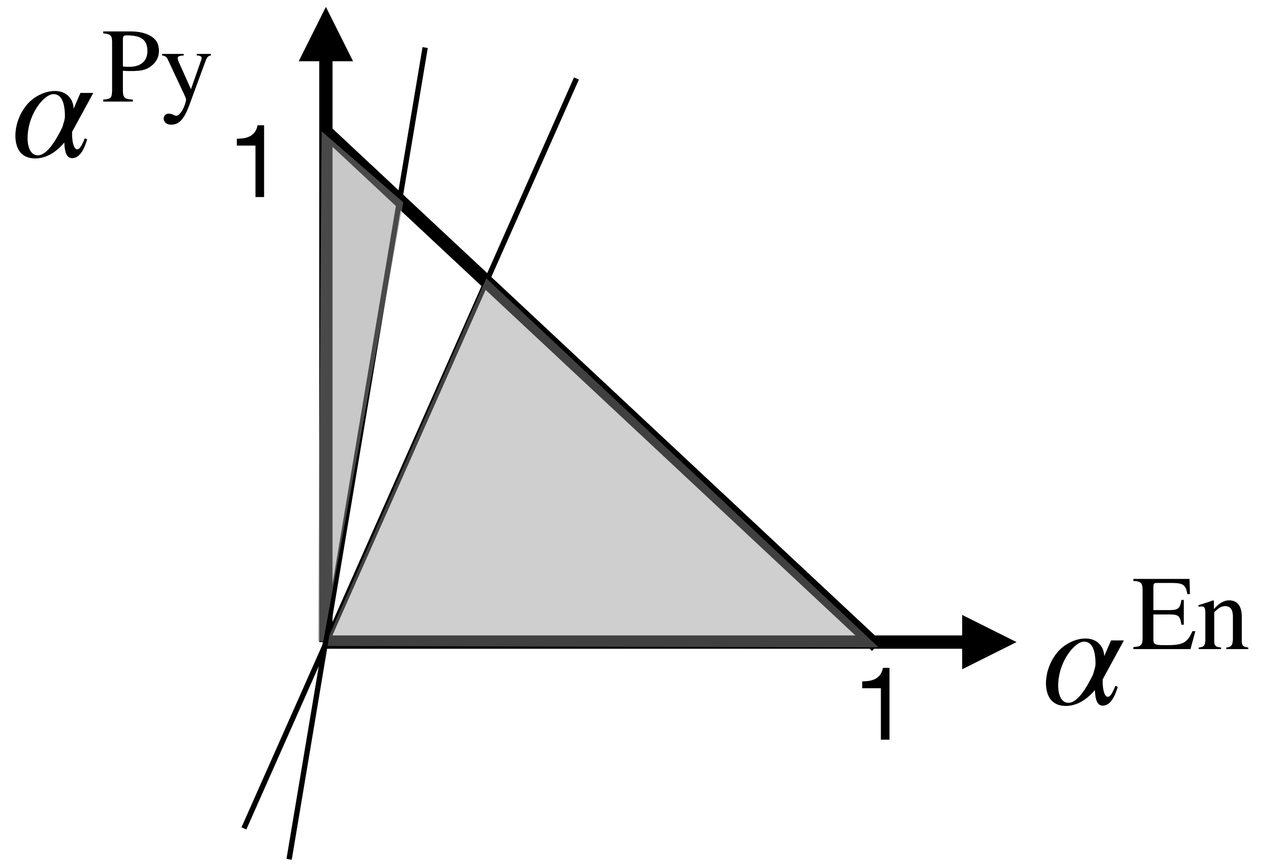
	_ _	( )	in	_t
En	6	1	1774	2246
Py	13055	248	1083	648

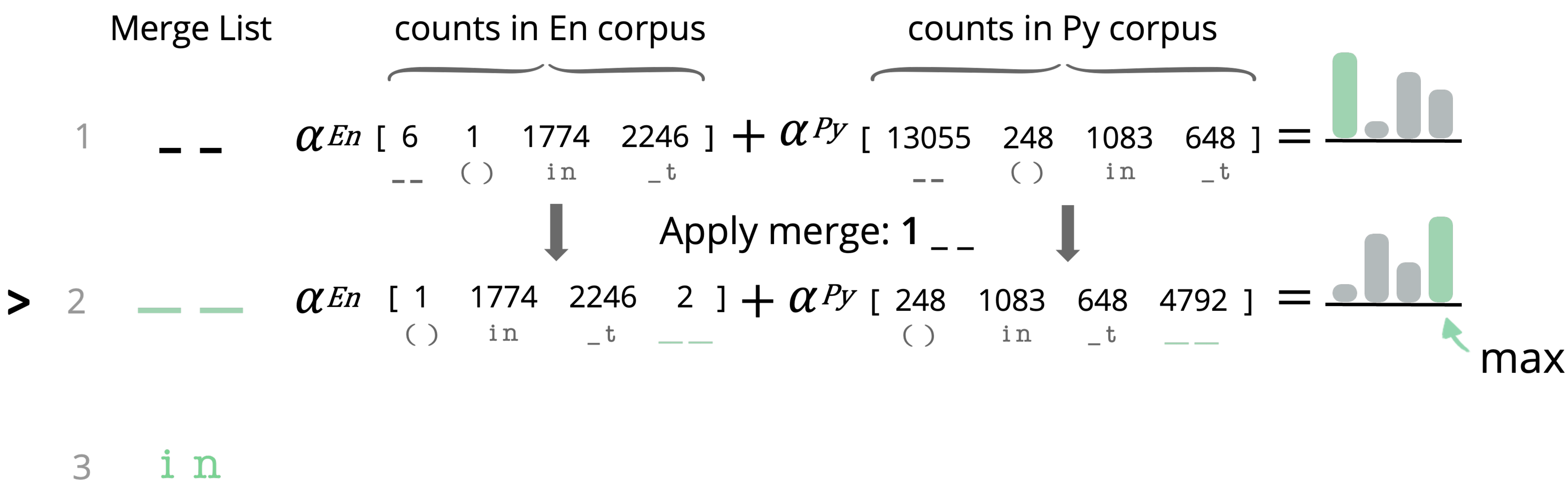
2 **_ _**

3 **i n**

Each token gives a specific linear condition that α_{En} and α_{Py} need to satisfy, for example:

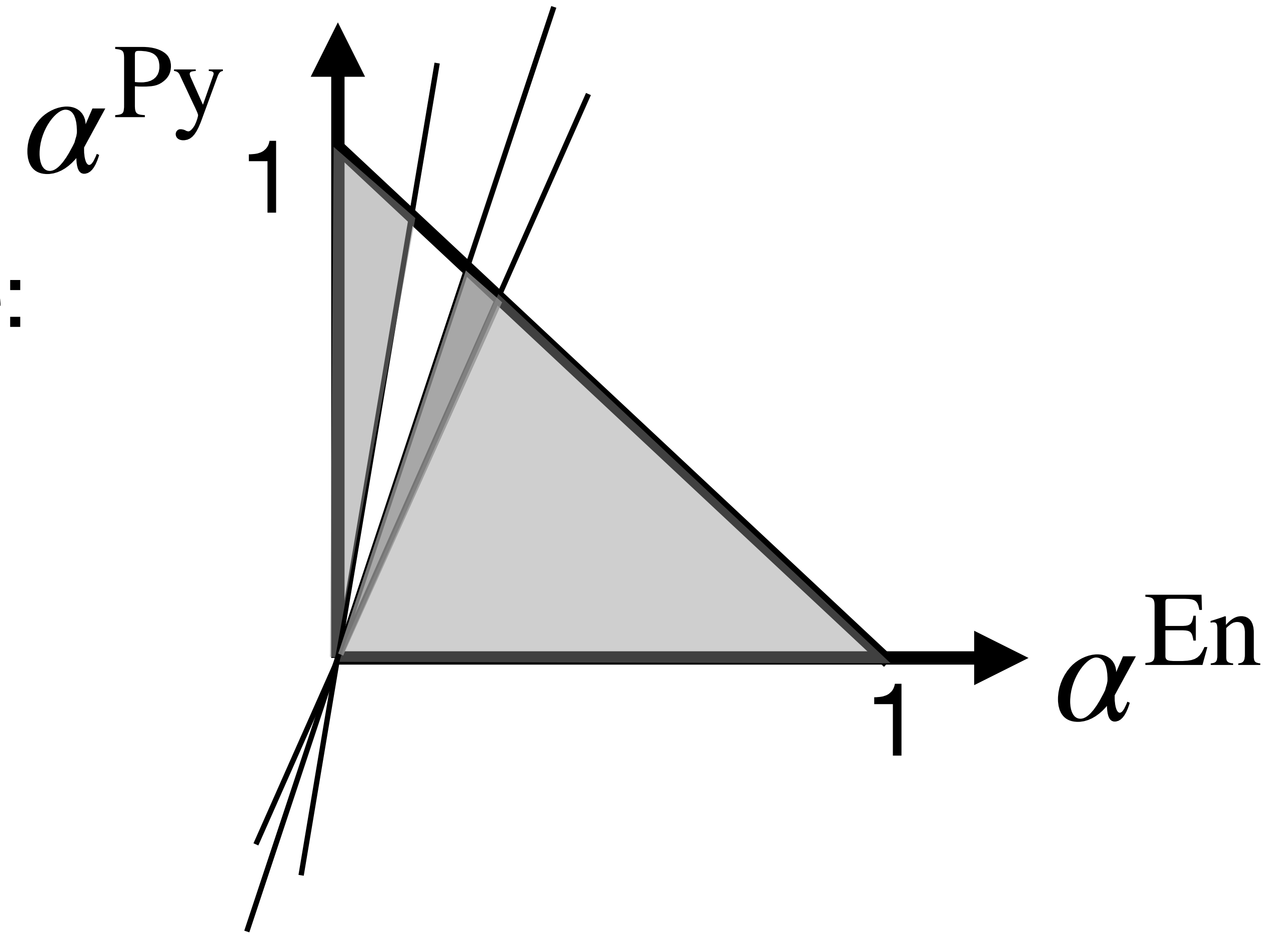
$$6\,\alpha_{\mathrm{En}} + 13055\,\alpha_{\mathrm{Py}} \geq \max_{token\,!=\_\_} \left\{ \alpha_{\mathrm{En}} C^{(1)}_{\mathrm{En},token} + \alpha_{\mathrm{Py}} C^{(1)}_{\mathrm{Py},token} \right\}$$

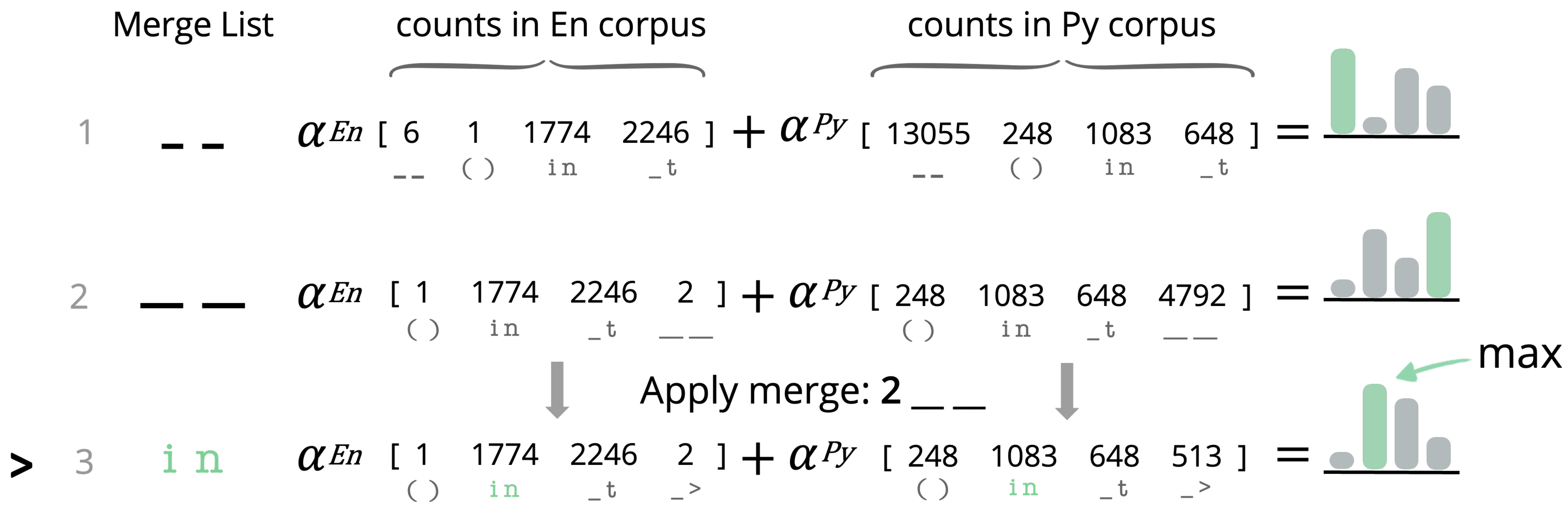




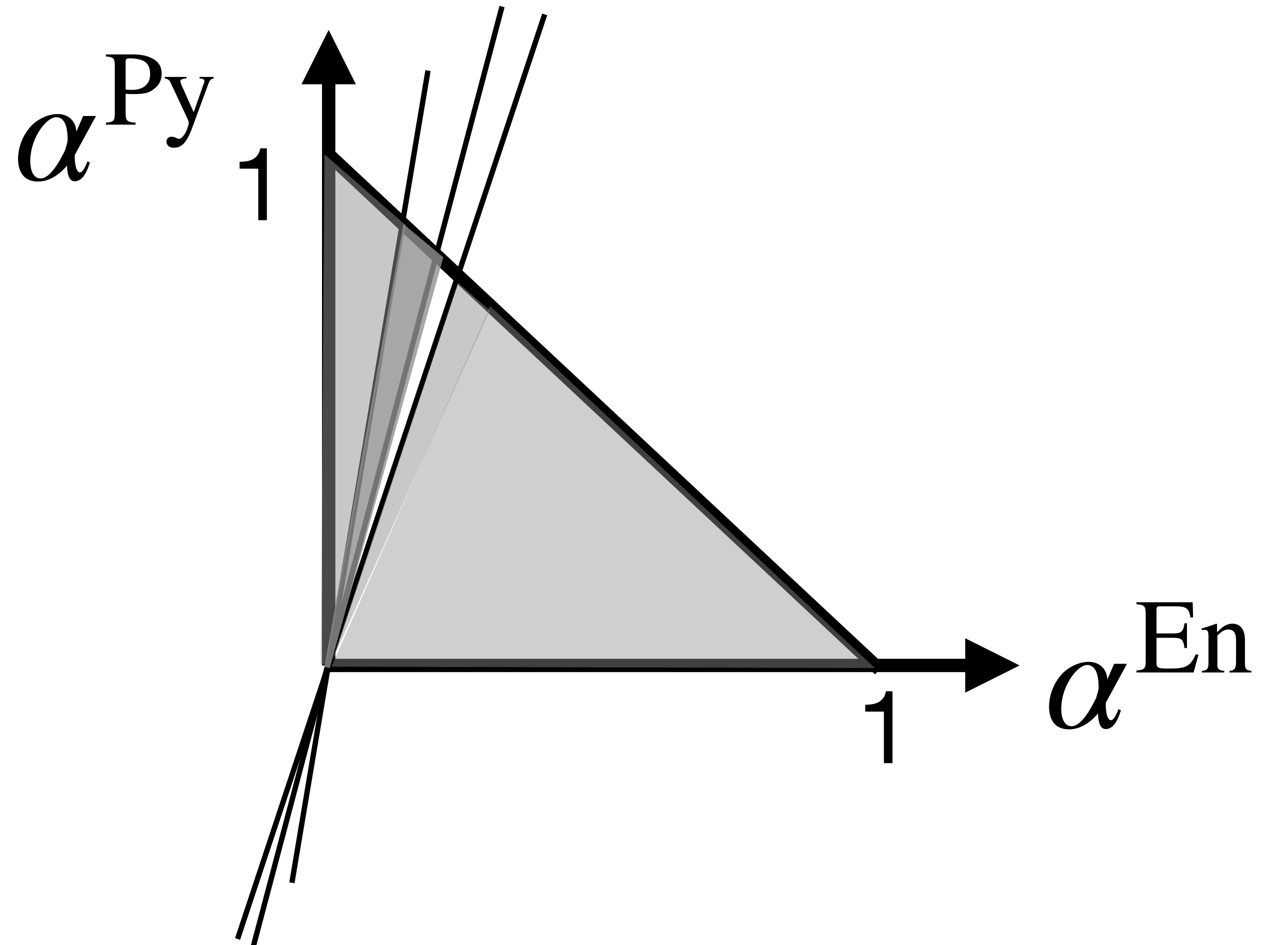
Each token gives a specific linear condition that α_{En} and α_{Py} need to satisfy, for example:

$$2\,\alpha_{\text{En}} + 4792\,\alpha_{\text{Py}} \geq \max_{token\,!=\,\_\_\,\_\_} \left\{ \alpha_{\text{En}} C^{(2)}_{\text{En},token} + \alpha_{\text{Py}} C^{(2)}_{\text{Py},token} \right\}$$

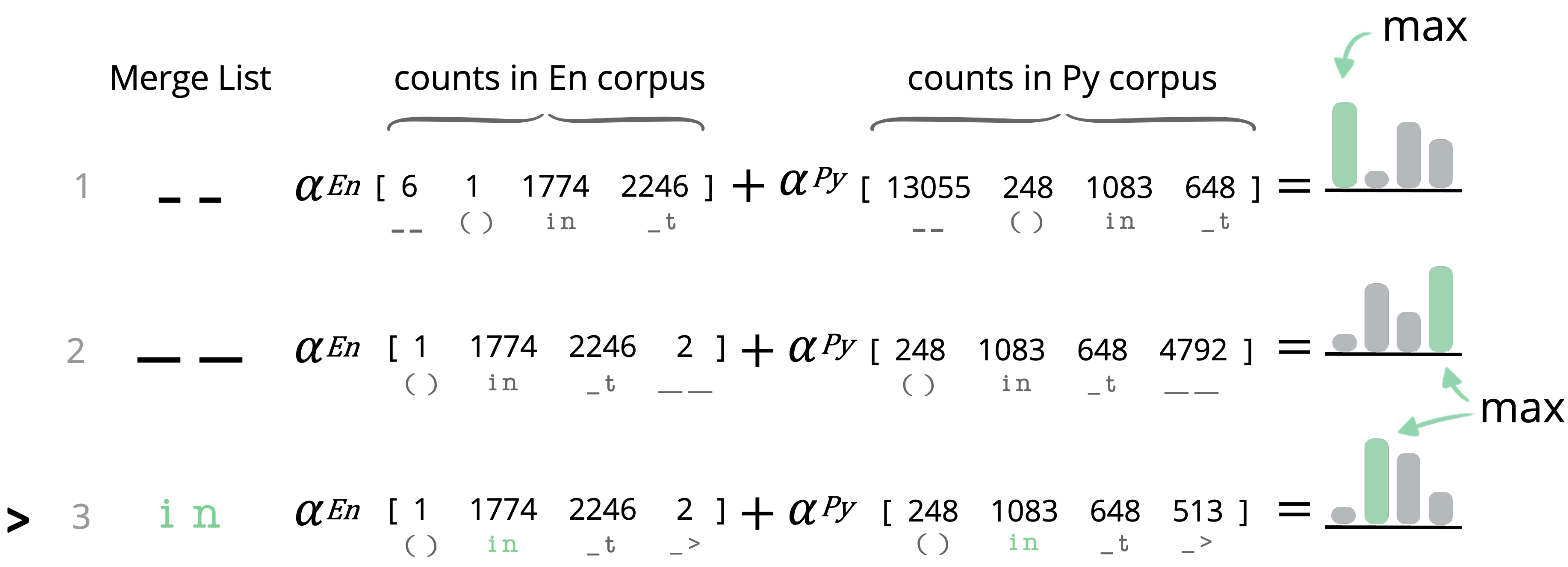




Each token gives a specific linear condition that α_{En} and α_{Py} need to satisfy, for example:



$$1774\ \alpha_{\text{En}} + 1083\ \alpha_{\text{Py}} \geq \max_{token\, != in} \left\{ \alpha_{\text{En}} C^{(3)}_{\text{En},token} + \alpha_{\text{Py}} C^{(3)}_{\text{Py},token} \right\}$$



At every step, the mixture ratios should give a vector with the true merge's index as the max value.

$$\sum_{i=1}^{n} \alpha_i c_{i,m^{(t)}}^{(t)} \geq \sum_{i=1}^{n} \alpha_i c_{i,p}^{(t)} \quad \text{for all } p \neq m^{(t)}$$

We can formulate this as a linear program

Objective: $\text{minimize} \sum_{t=1}^{M} v^{(t)} + \sum_{p} v_p$

Subject to constraints:

At every time step t,

constraint violation

$$v^{(t)} + v_p + \sum_{i=1}^{n} \alpha_i c_{i,m^{(t)}}^{(t)} \geq \sum_{i=1}^{n} \alpha_i c_{i,p}^{(t)} \quad \text{for all } p \neq m^{(t)}$$

for each time step t

for each pair p

Controlled Experiments

Evaluate attack on tokenizers trained with known mixtures!

- **Natural languages** (112) from Oscar (web data)
- **Programming languages** (37) from raw Github data
- **Domains** (5) from RedPajama (all English) — web, books, Wiki, code, ArXiv

For $n \in \{5,10,30,112\}$, sample n categories and weights uniformly.

Sample 10G of data with the desired mixture ratio for tokenizer training. For the attack, sample 1G of data per category.

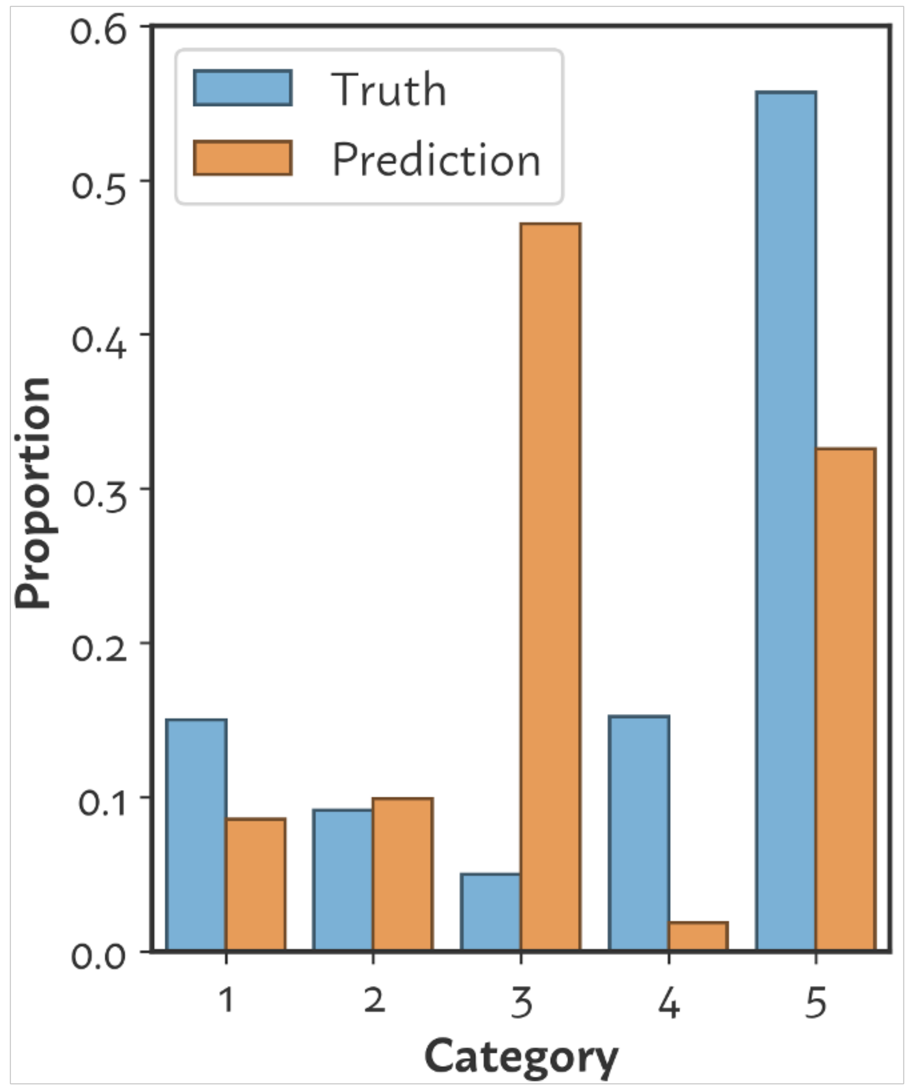
Report MSE $= \frac{1}{n}\sum_{i=1}^{n}(\hat{\alpha}_i - \alpha_i)^2$.

Results

Log$_{10}$ MSE (↓)

n	Random	Languages	Code	Domains
5				
10				
30				
112				

number of categories

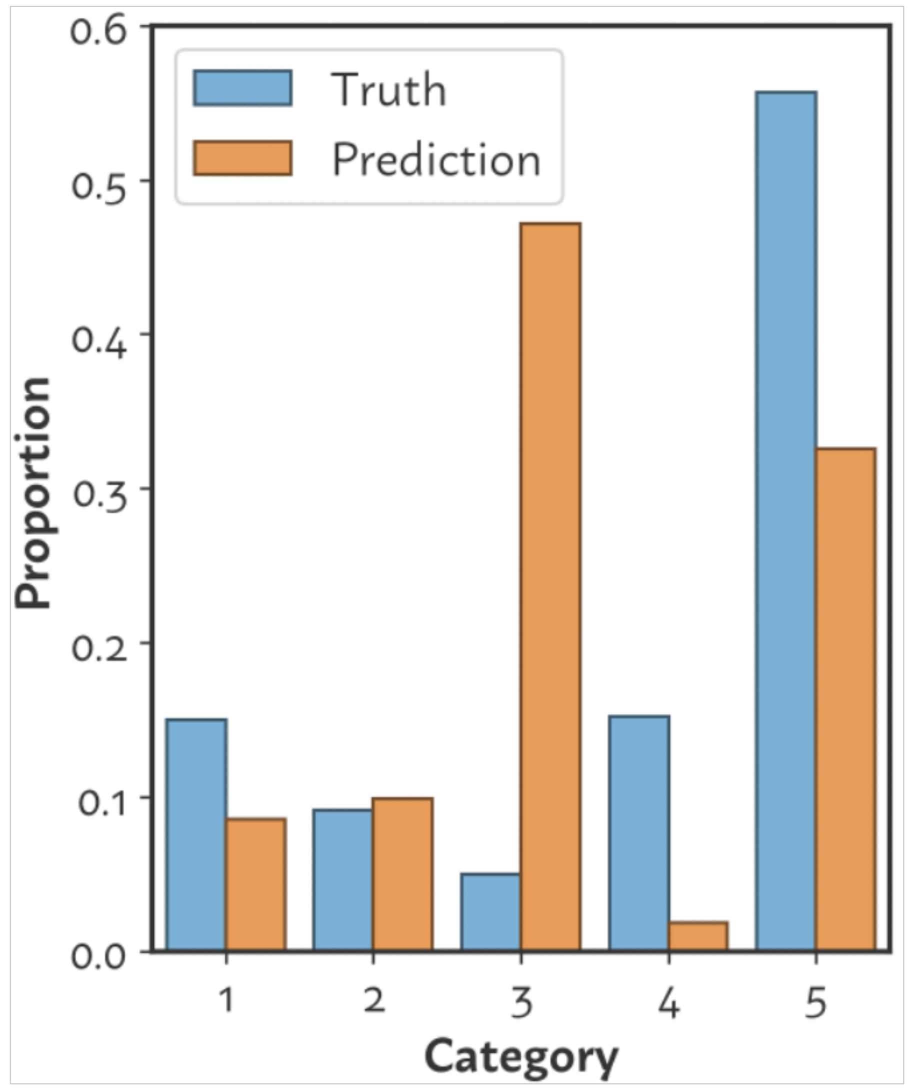


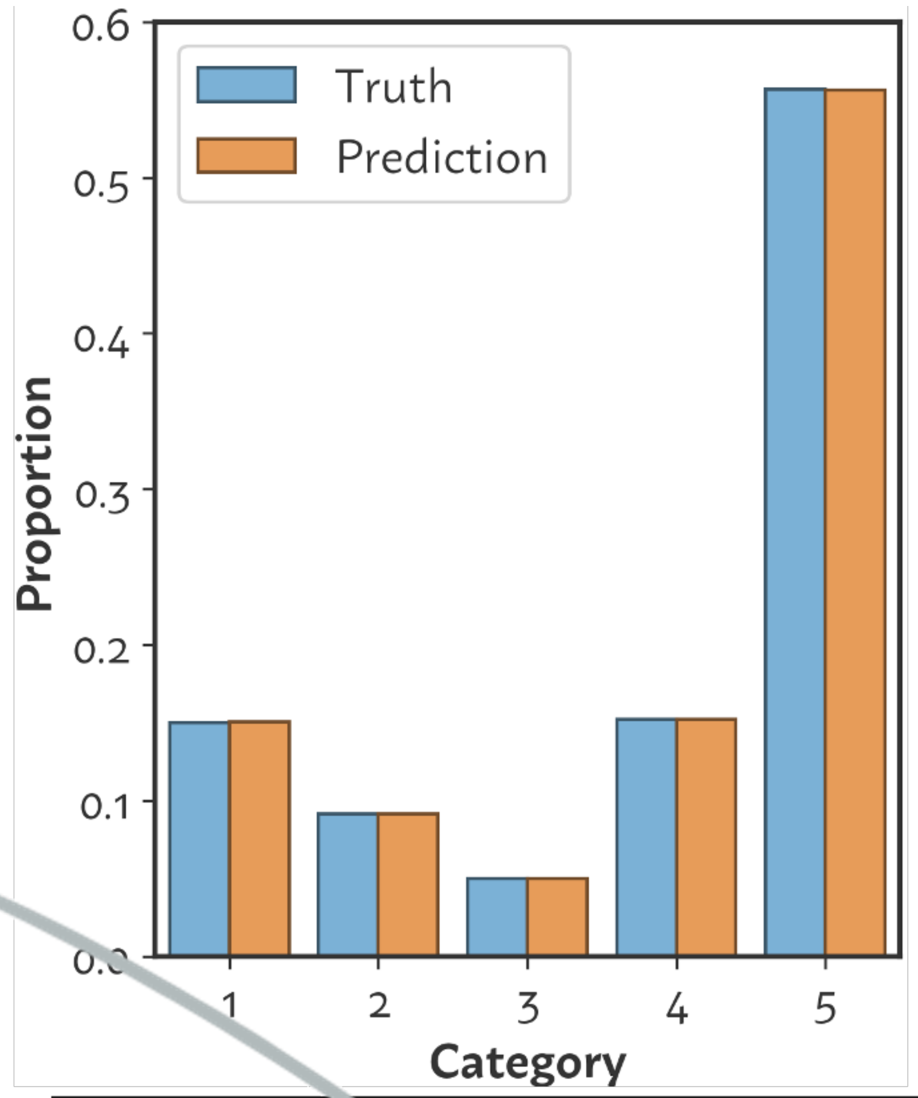
Results

Log_{10} MSE (↓)

n	random guess baseline	Languages	Code	Domains
5	-1.39			
10				
30				
112				

number of categories



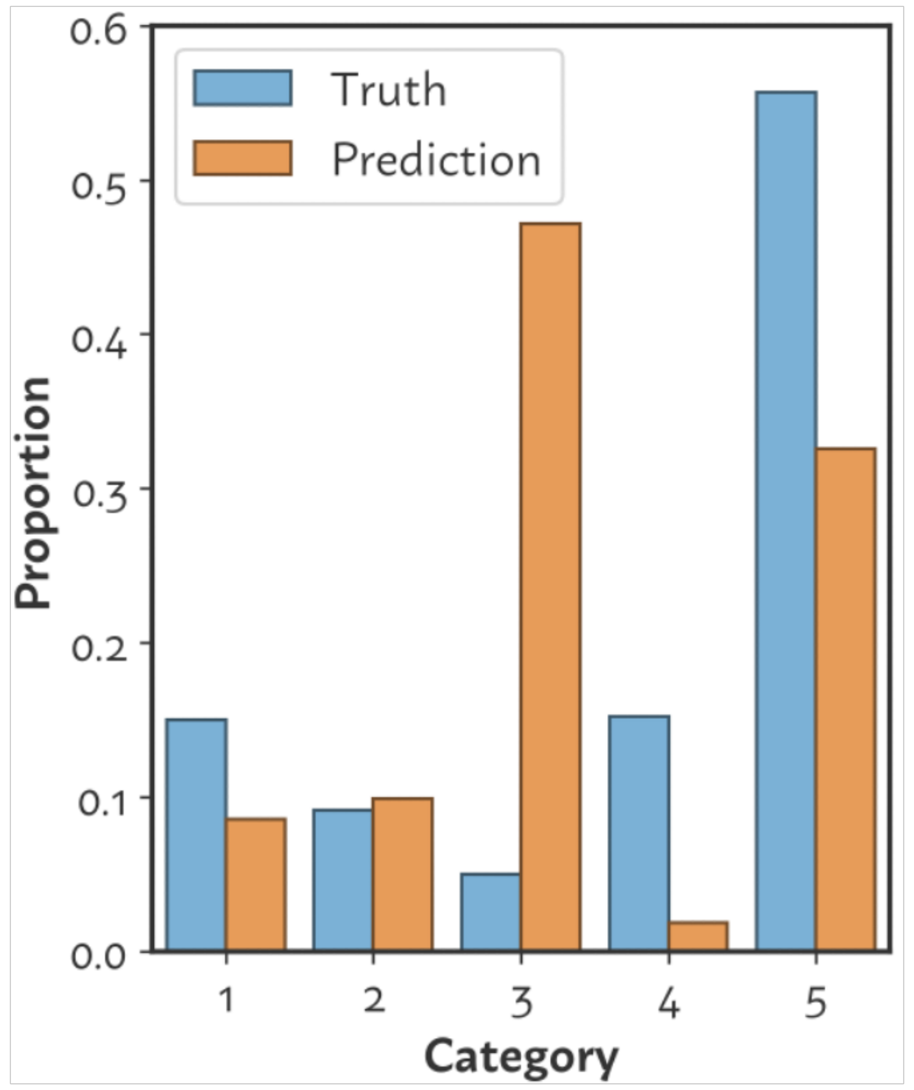


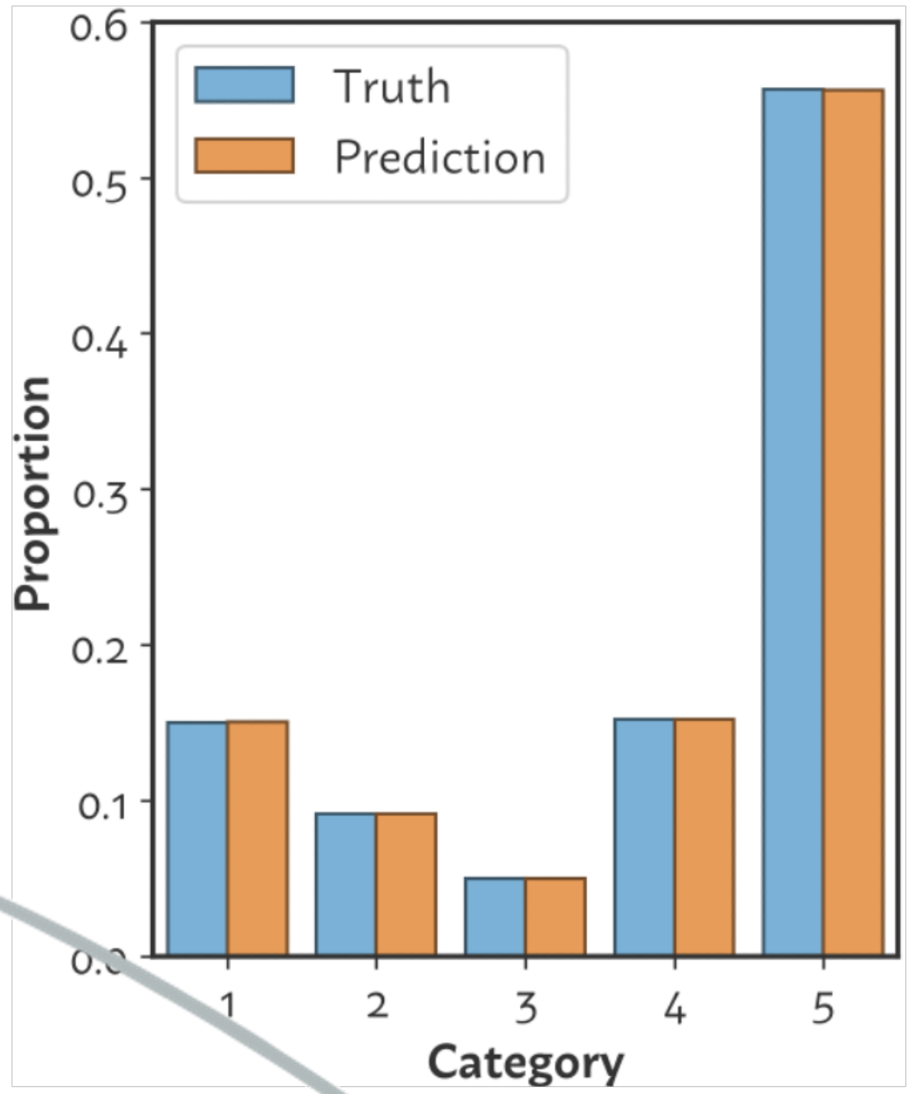
Results

Log$_{10}$ MSE (↓)

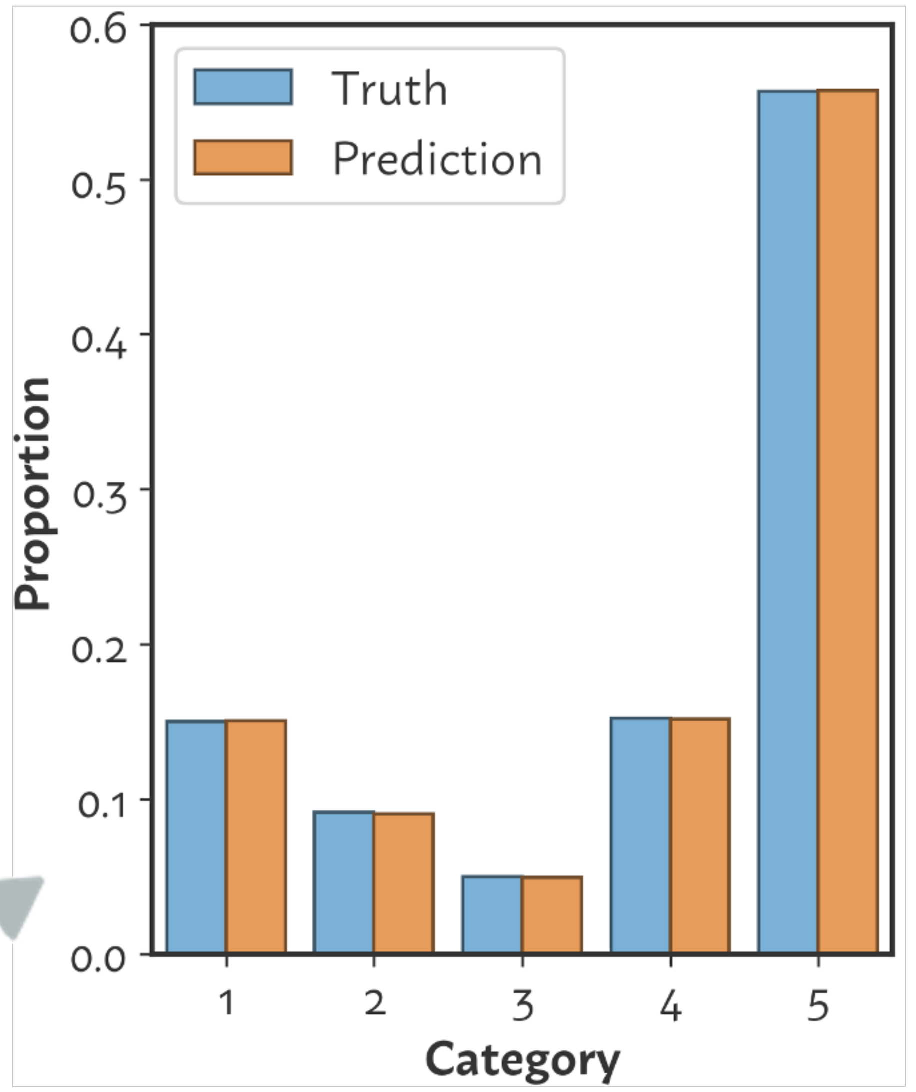
number of categories

n	Random	Languages	Code	Domains
5	-1.39	-7.30		
10				
30				
112				





Results

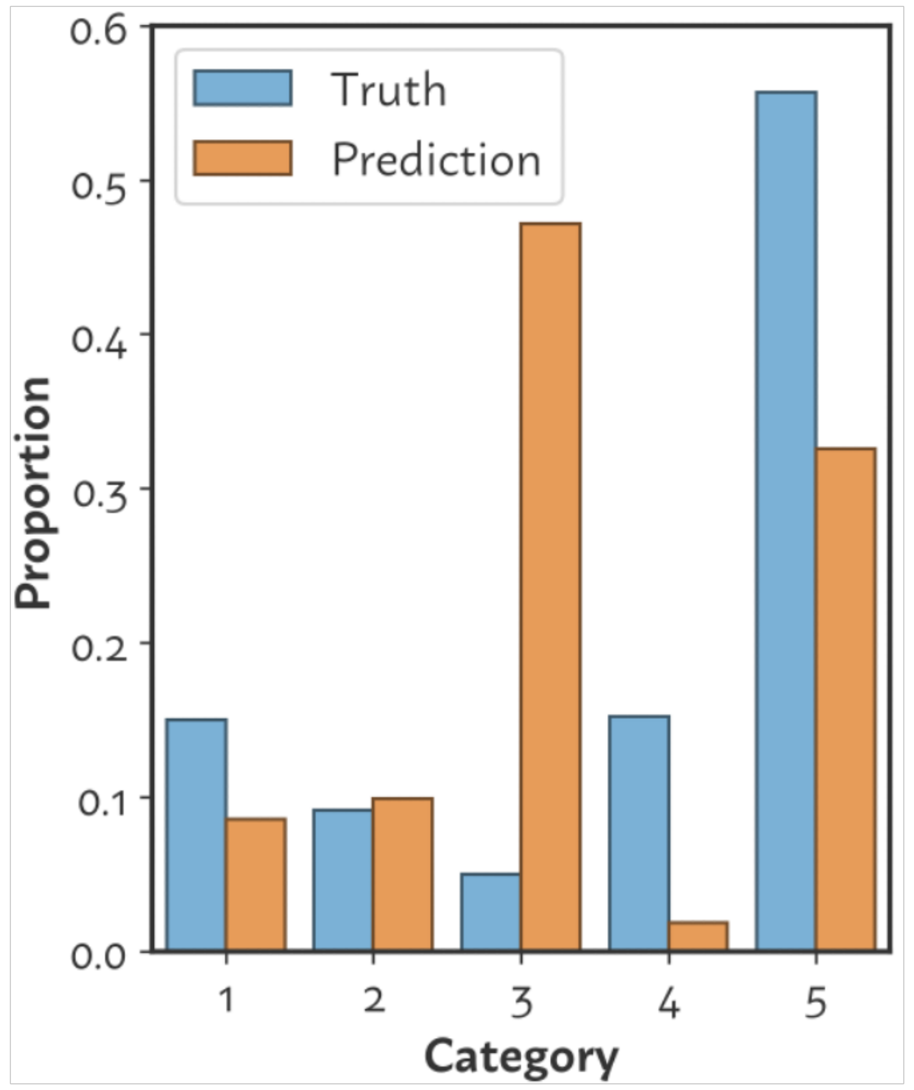


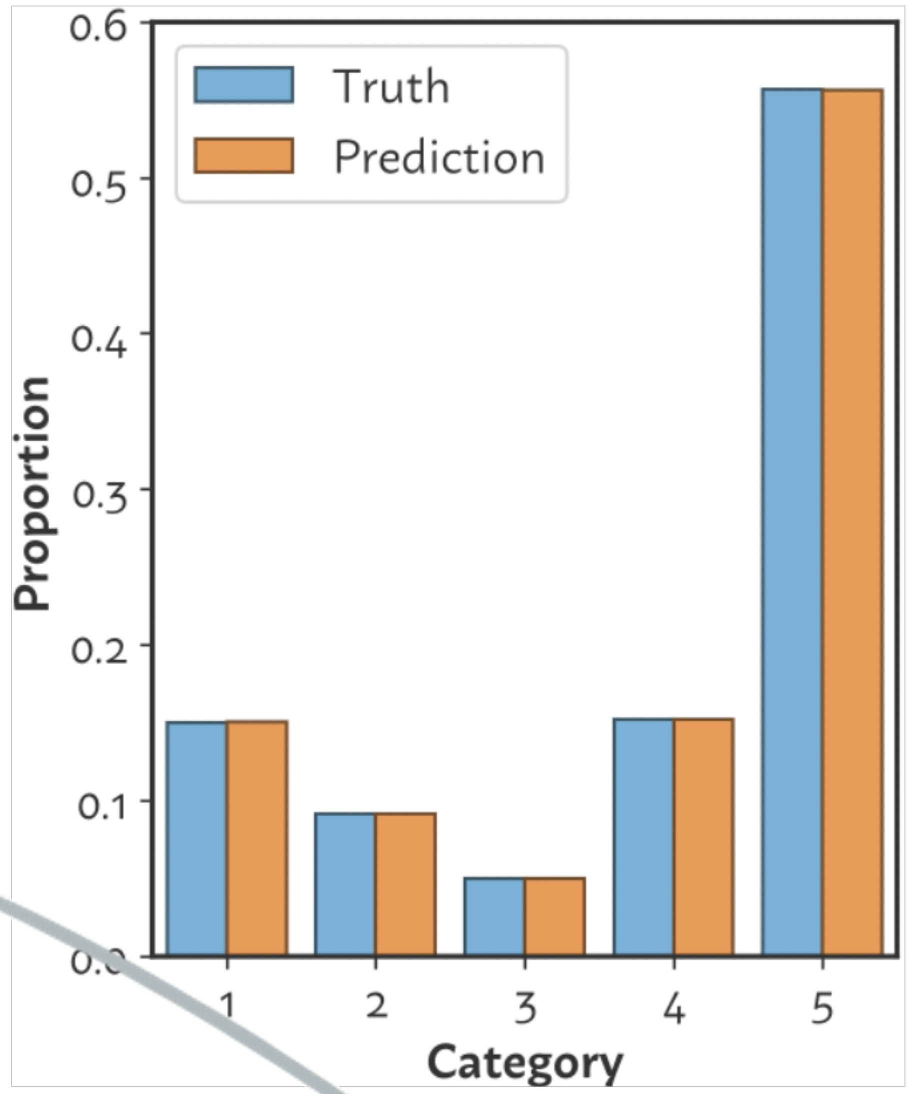
Log_{10} MSE (↓)

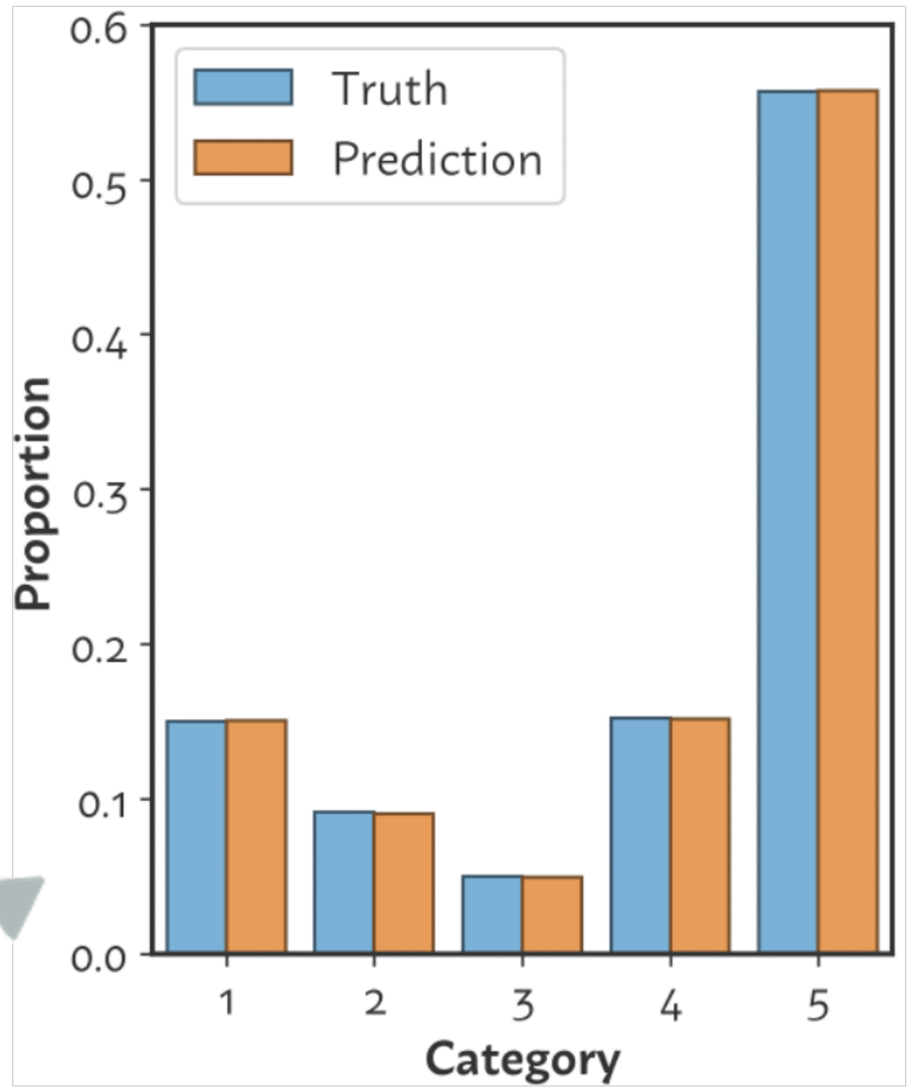
n	Random	Languages	Code	Domains
5	-1.39	-7.30	-6.46	
10				
30				
112				

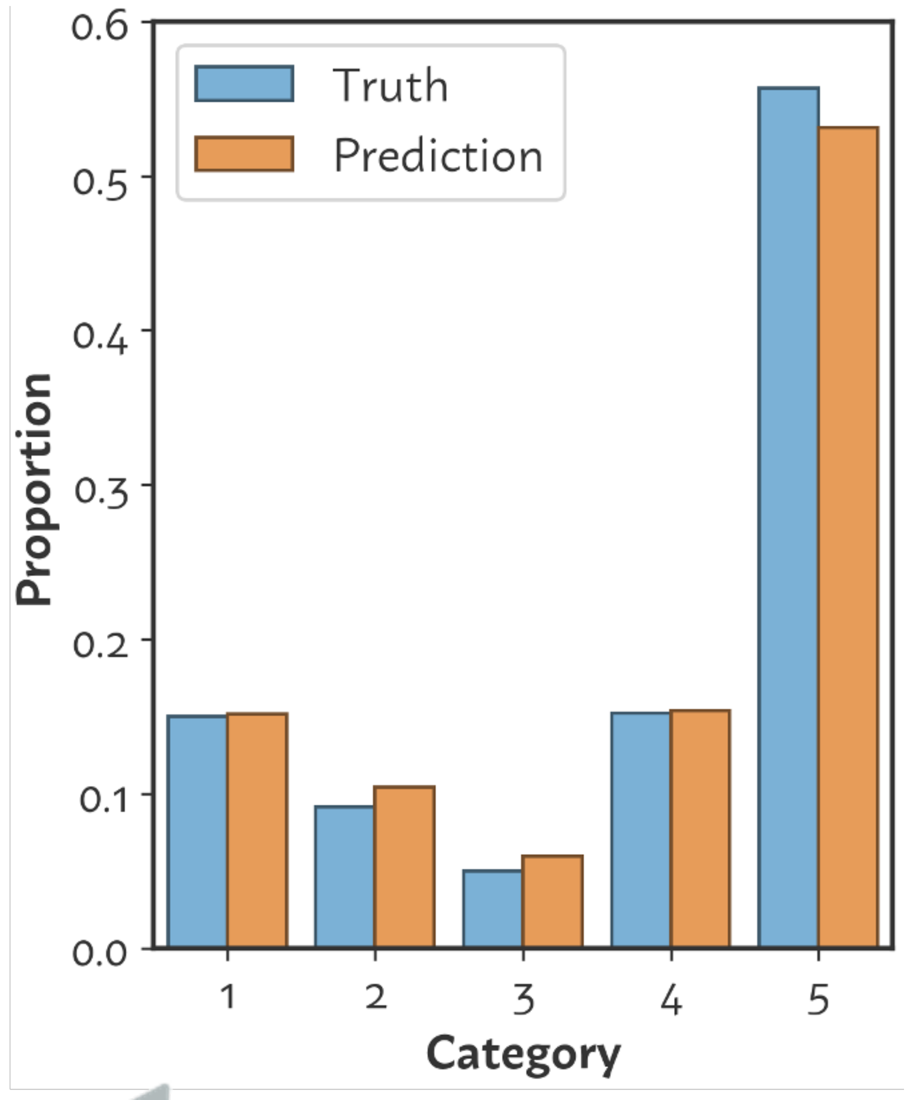
number of categories

Results









$\log_{10}$ MSE (↓)

n	Random	Languages	Code	Domains
5	-1.39	-7.30	-6.46	-3.74
10				
30				
112				

number of categories

Log$_{10}$ MSE (↓)

number of categories

n	Random	Languages	Code	Domains
5	-1.39	-7.30	-6.46	-3.74
10	-1.84	-7.66	-6.30	-
30	-2.70	-7.73	-5.98	-
112	-3.82	-7.69	-	-

Our attack achieves performance 10^2 to $10^6\times$ better than random!

Commercial Tokenizers

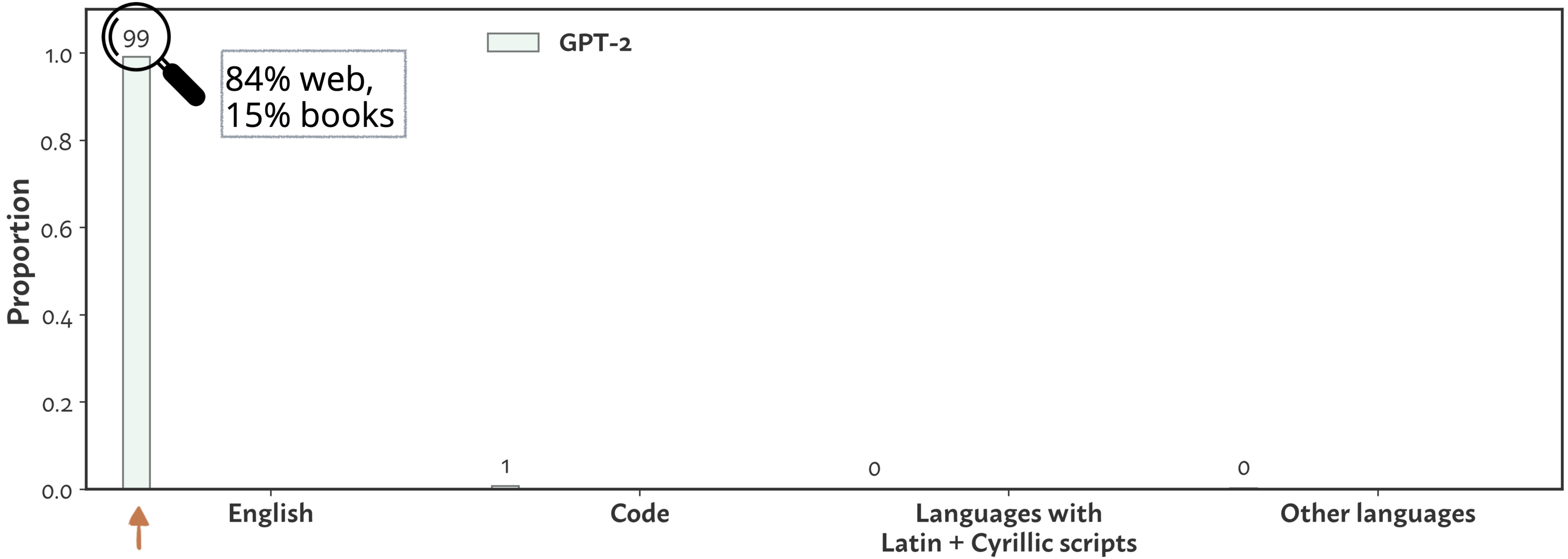
Let’s apply our attack to off-the-shelf tokenizers released with LLMs!

Total set of 116 categories: 111 languages, code, and 4 En domains.

- Split “English” into 4 En domains: web, Wikipedia, ArXiv, books.
- Combine programming languages into 1 code domain.

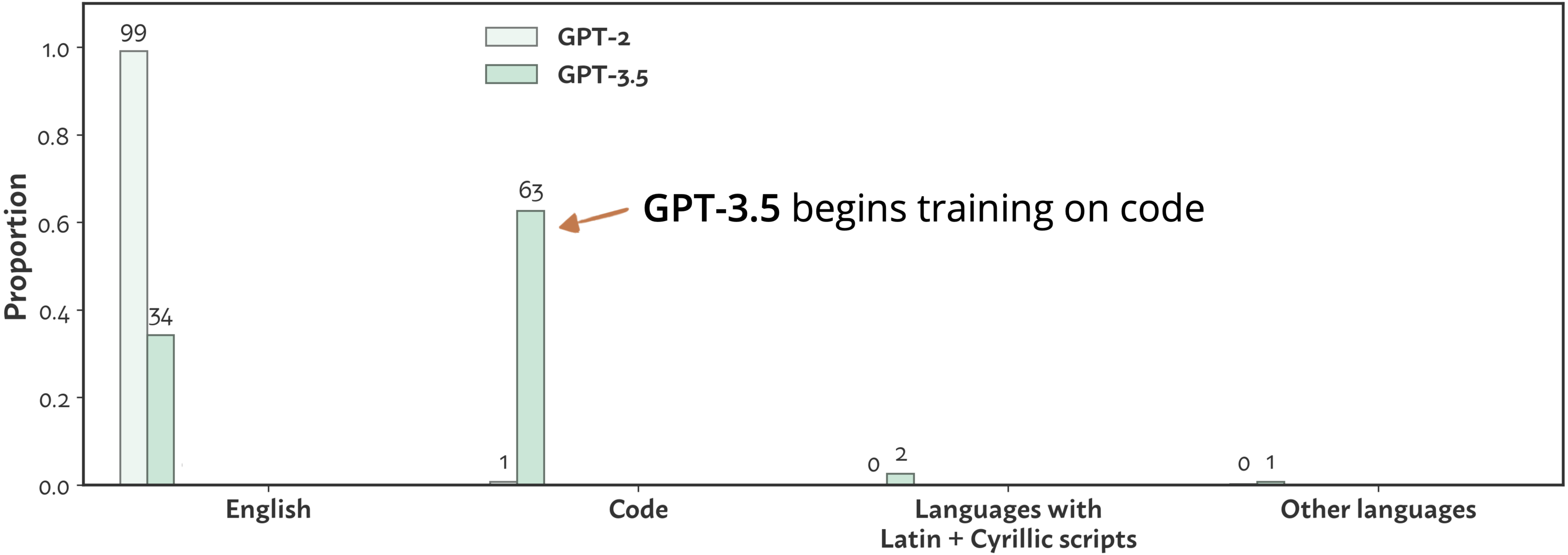
We study: GPT-2, GPT-3.5, GPT-4o, Llama, Llama 3, Mistral, Mistral-Nemo, GPT-NeoX, Gemma, Claude, Command R, …

Our Inference for LLM Tokenizers

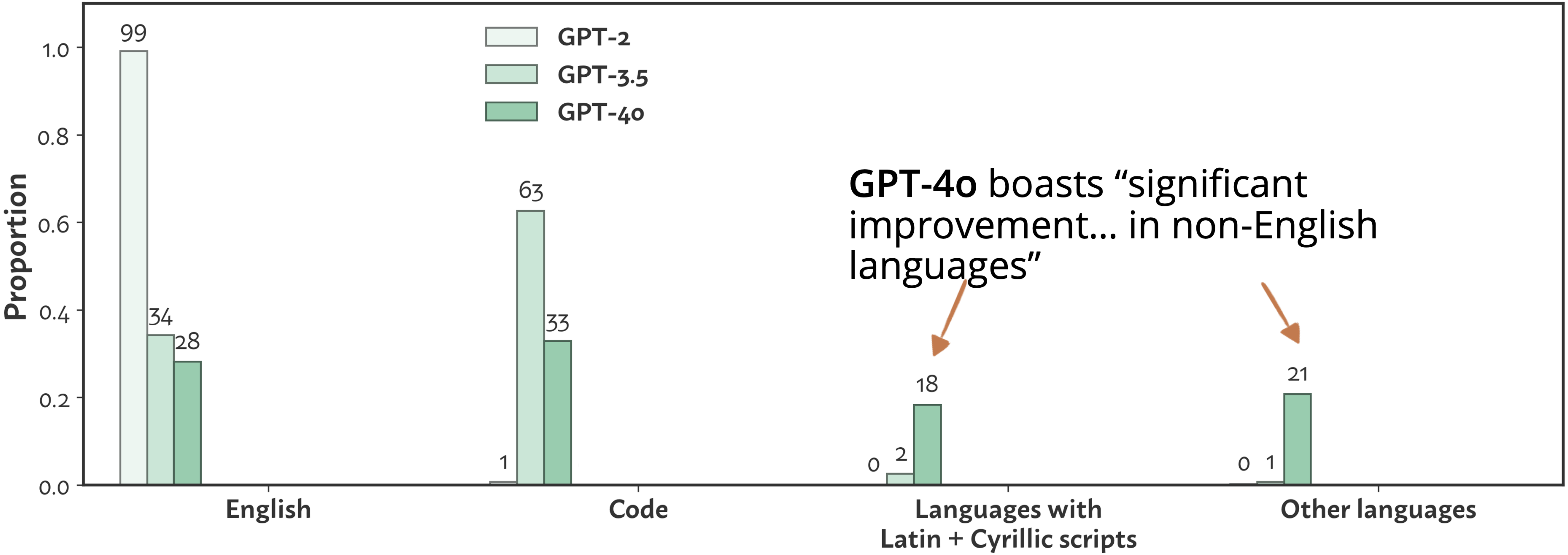


For **GPT-2**, “a filter was used to produce an English only dataset”

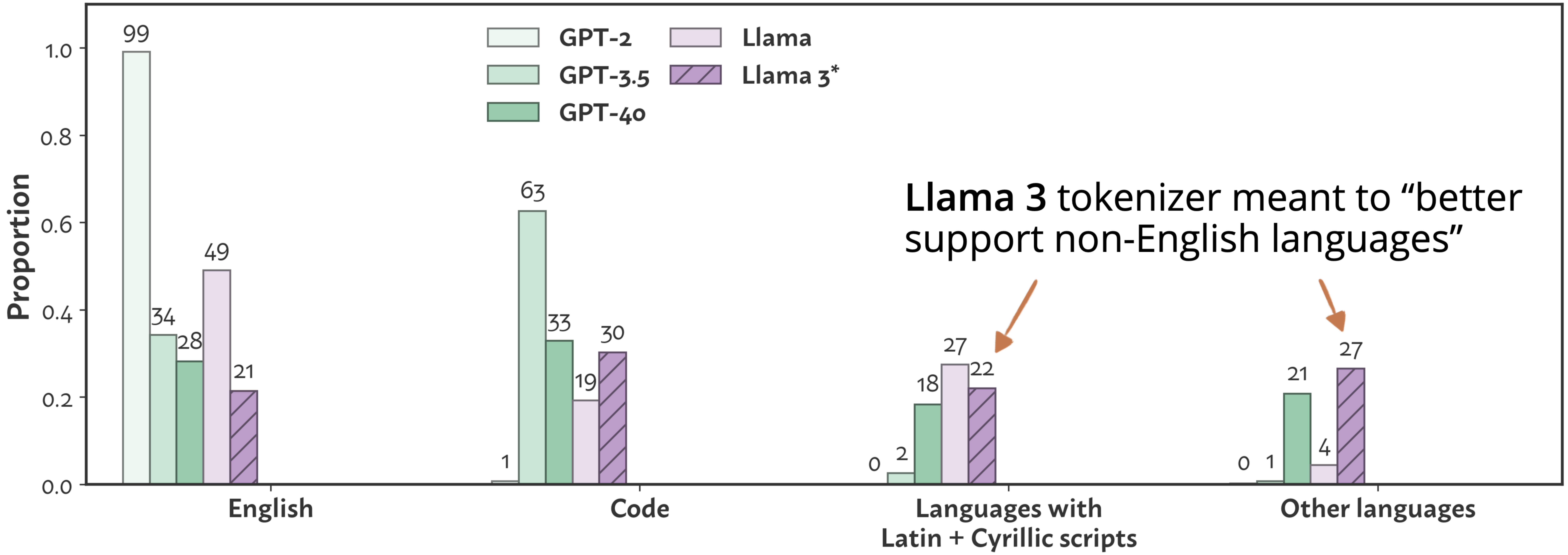
Our Inference for LLM Tokenizers



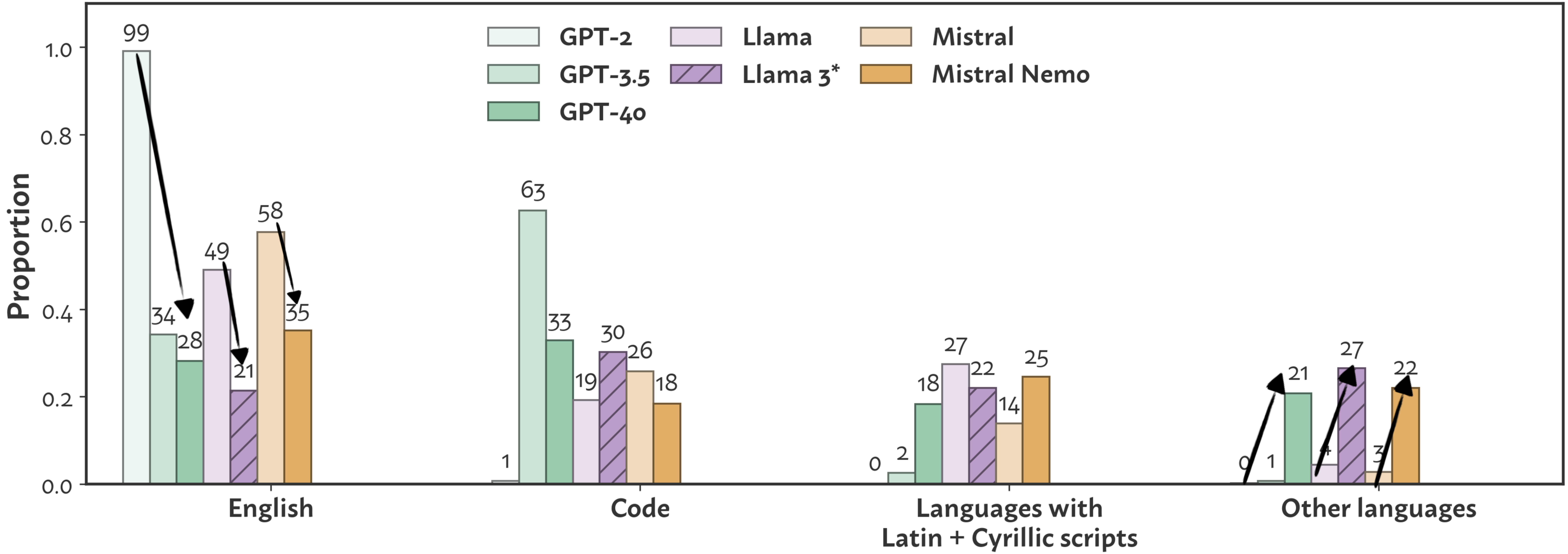
Our Inference for LLM Tokenizers



Our Inference for LLM Tokenizers

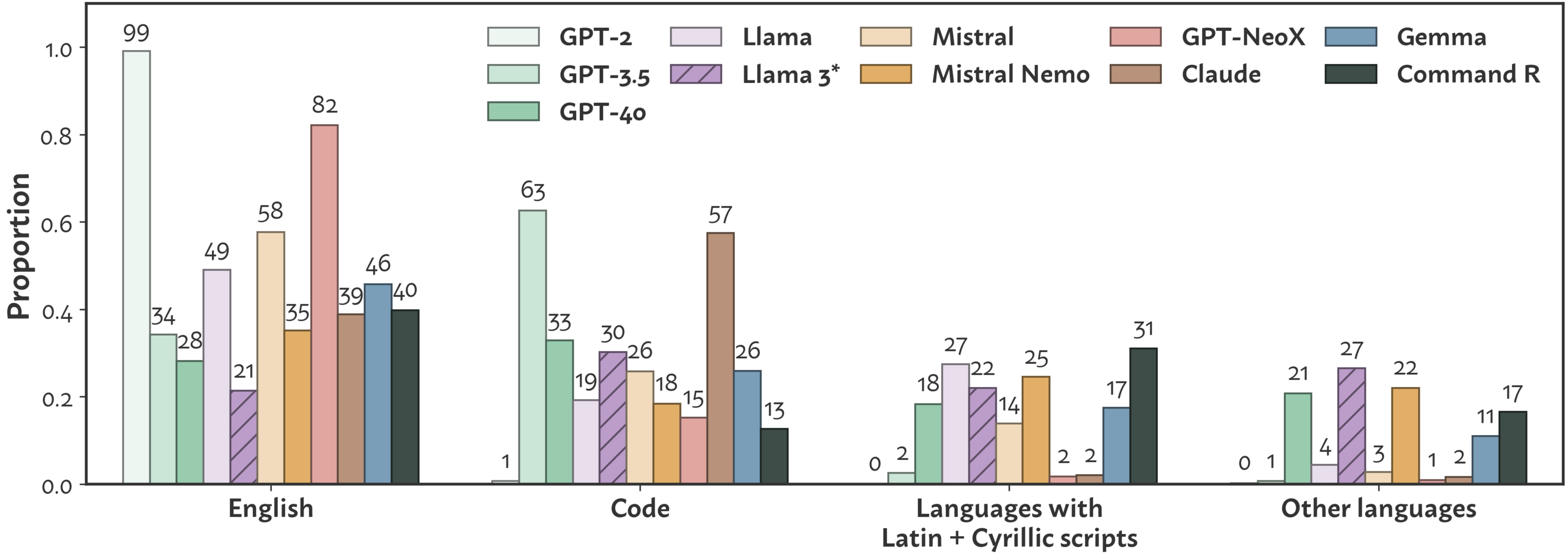


Our Inference for LLM Tokenizers



Trend: newer generations of models are more multilingual

Our Inference for LLM Tokenizers



Takeaways

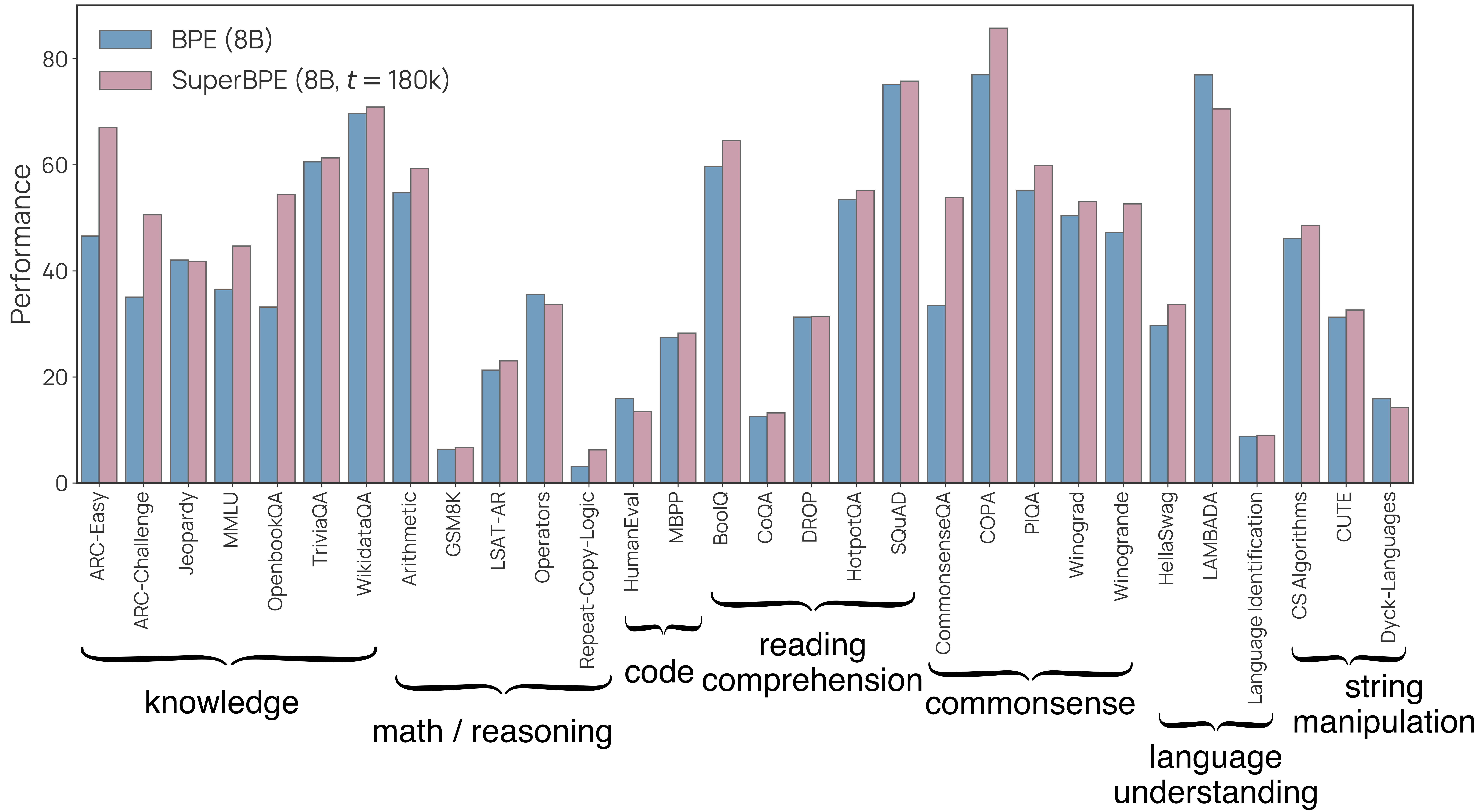
- Pretraining dataset is a trade secret
- Dataset mixture inference from BPE tokenizer reliably recovers the mixture weights, allowing us to peak into what choices were made in the evolution of language models

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,

- **“Data Mixture Inference Attack: BPE Tokenizers Reveal Training Data Compositions”**, Jonathan Hayase, Alisa Liu, Yejin Choi, Sewoong Oh, Noah A. Smith, *NeurIPS 2024*

SuperBPE downstream performance



Efficiency scaling for non-English languages