

PAUL G.
ALLEN
SCHOOL



nVIDIA®



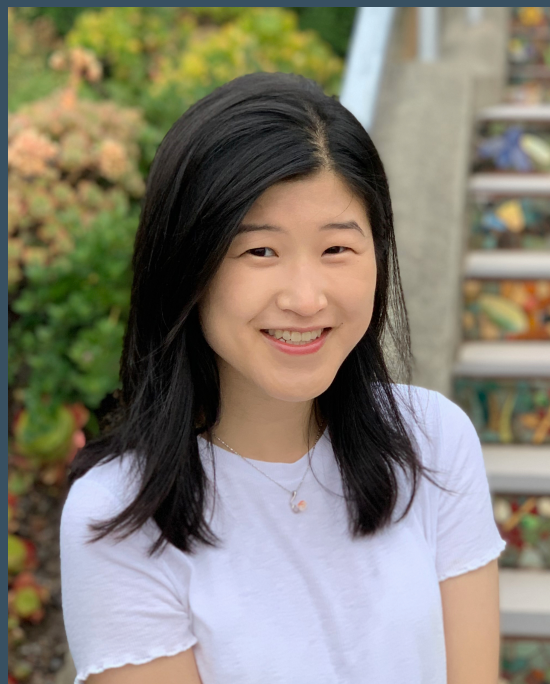
SuperBPE: 

Space Travel for Language Models

Microsoft Translator

Apr 16, 2025

* Alisa Liu



* Jonathan Hayase



Valentin
Hofmann



Sewoong Oh Noah Smith Yejin Choi

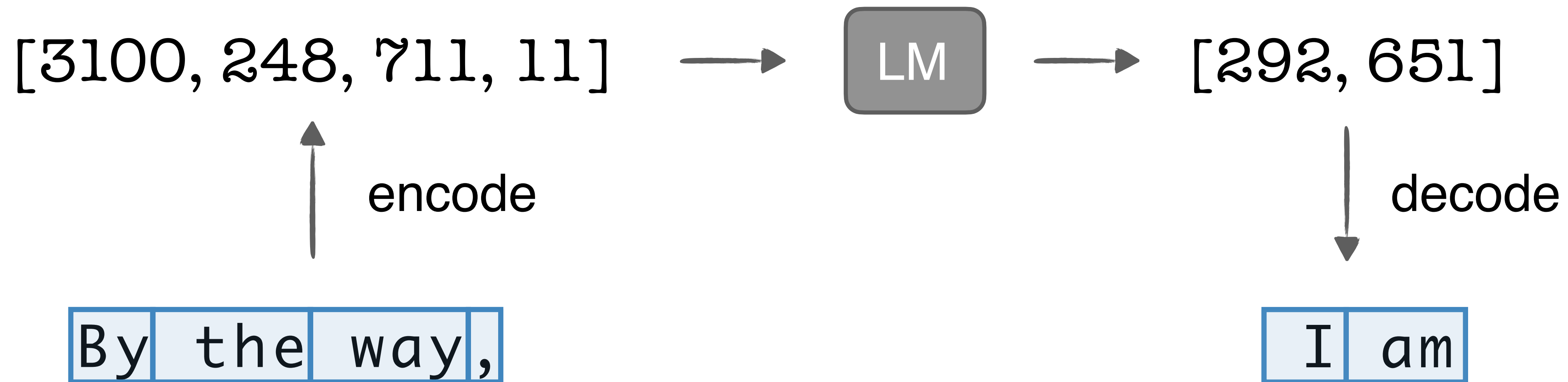


Tokenization controls representation of text

Token = basic textual unit with its own embedding representation

The **tokenizer** encodes text into a sequence of **tokens** (and back)

Directly controls the representation of text that LMs learn over!



Tokenization today: subword level

Transformer LMs (almost) universally use **subwords** as tokens

Character-level: can represent all text, but extremely inefficient

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

Word-level: more efficient, but requires UNK

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

Subword-level: best of both worlds!

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

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

Whitespace is not a reliable delimiter of meaning!

Multi-word expressions

“by the way,” “by accident,” “for a living,” “in the long run”

Crosslingually, the same meaning can be 1 word or several words

“math teacher” → “Mathelehrer” in German

“your friend” → “arkadaşın” in Turkish

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)

(Near) universal method today for learning subword tokenizers

Intuition: build the vocabulary bottom-up by repeatedly merging common token sequences into new tokens

Introduced by [Sennrich et al., 2016](#) & popularized by [GPT-2 \(2019\)](#)

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

_ 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 i n 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
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 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

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 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

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 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

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 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

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 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

_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 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

_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 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

_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 i n g, _ o f, _ m a
n y, _ s t a r s, _ c a m
e, _ in, _ 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, _the, _ 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 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 i n g, _ o f, _ m a
n y, _ s t a r s, _ c a m
e, _ in, _ 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, _the, _ 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

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 i n g, _ o f, _ m a
n y, _ s t a r s, _ c a m
e, _ in, _ 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, _the, _ 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

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*

Modify BPE to create SuperBPE!

In BPE, whitespace pretokenization prevents merges across whitespace

SuperBPE introduces a *pretokenization curriculum*

Stage 1: Learn *subwords* using whitespace pretokenization, up until transition point $t < T$

Stage 2: Learn *superwords* by resuming training without whitespace pretokenization, until final vocabulary size T

Intuitively, learn basic units of meaning (words) then learn common word sequences (superwords!)

Getting the Most out of your Tokenizer (Dagan et al., 2024)

Tokenization is More than Compression (Schmidt et al., 2024)

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_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.}
```

Skip whitespace pretokenization
(but can still use other pretok rules!)

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

Tokenizer training setup

Train baseline BPE tokenizer and SuperBPE tokenizers

Vocab size: 200,000

Follows the recommendation of *Scaling Laws with Vocabulary (Tao et al., 2024)*

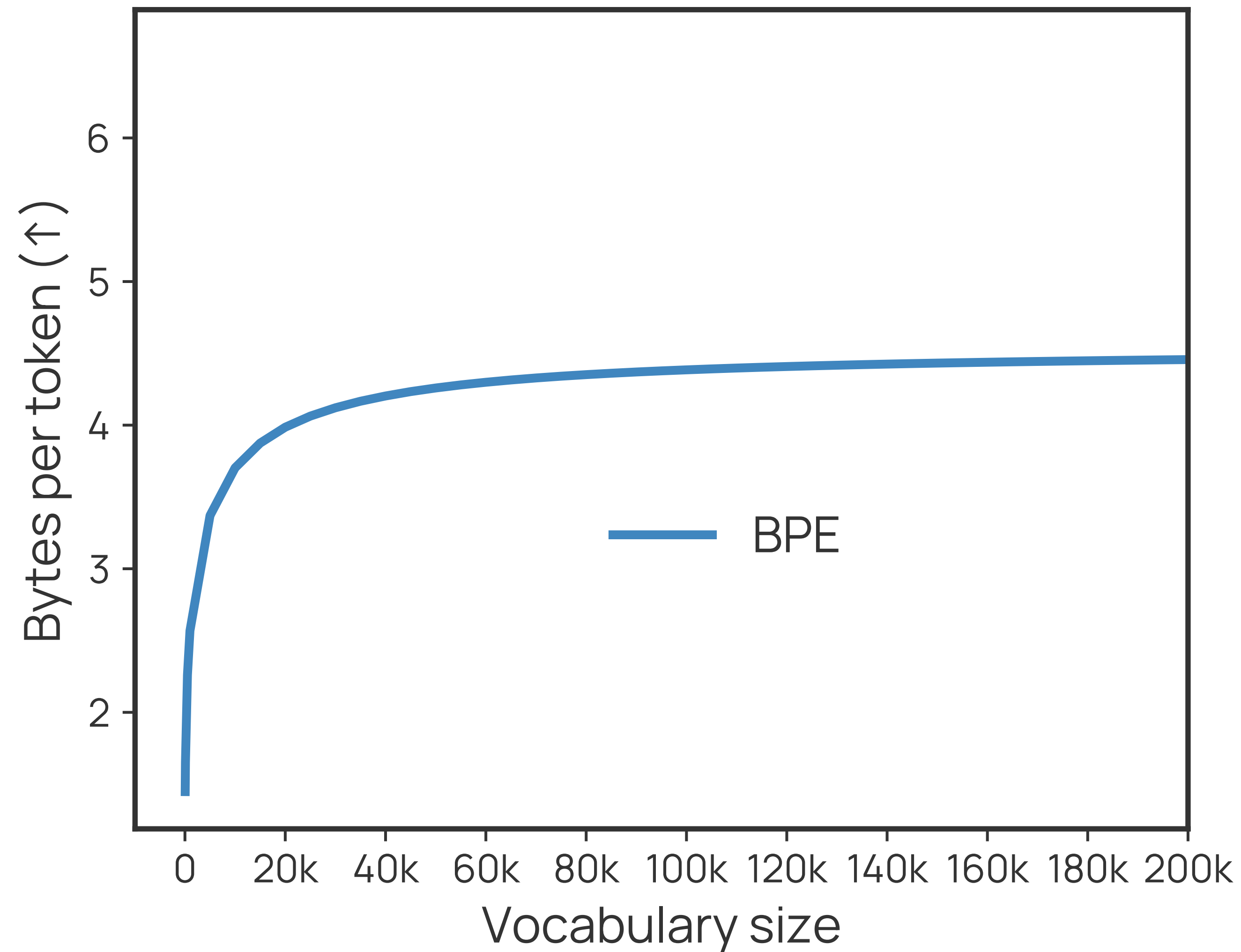
Transition points: {20K, 40K, ..., 180K}

Training data: 10 GB subset of Olmo 2 pretrain mix.

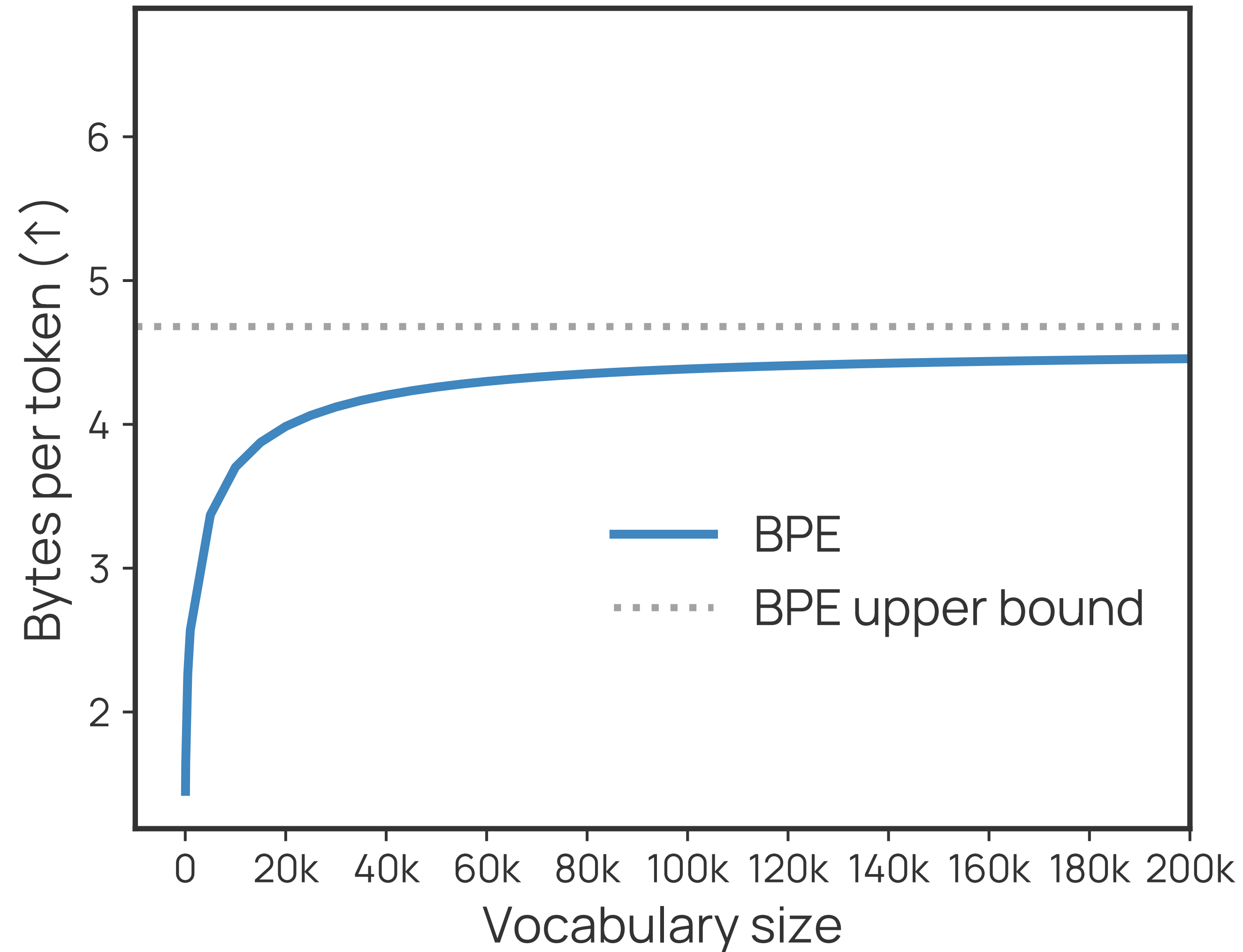
Pretokenization of numbers: groups of 3, right aligned ($1000 \rightarrow [1,000]$)

Limit superword tokens to 4 words to minimize privacy risk

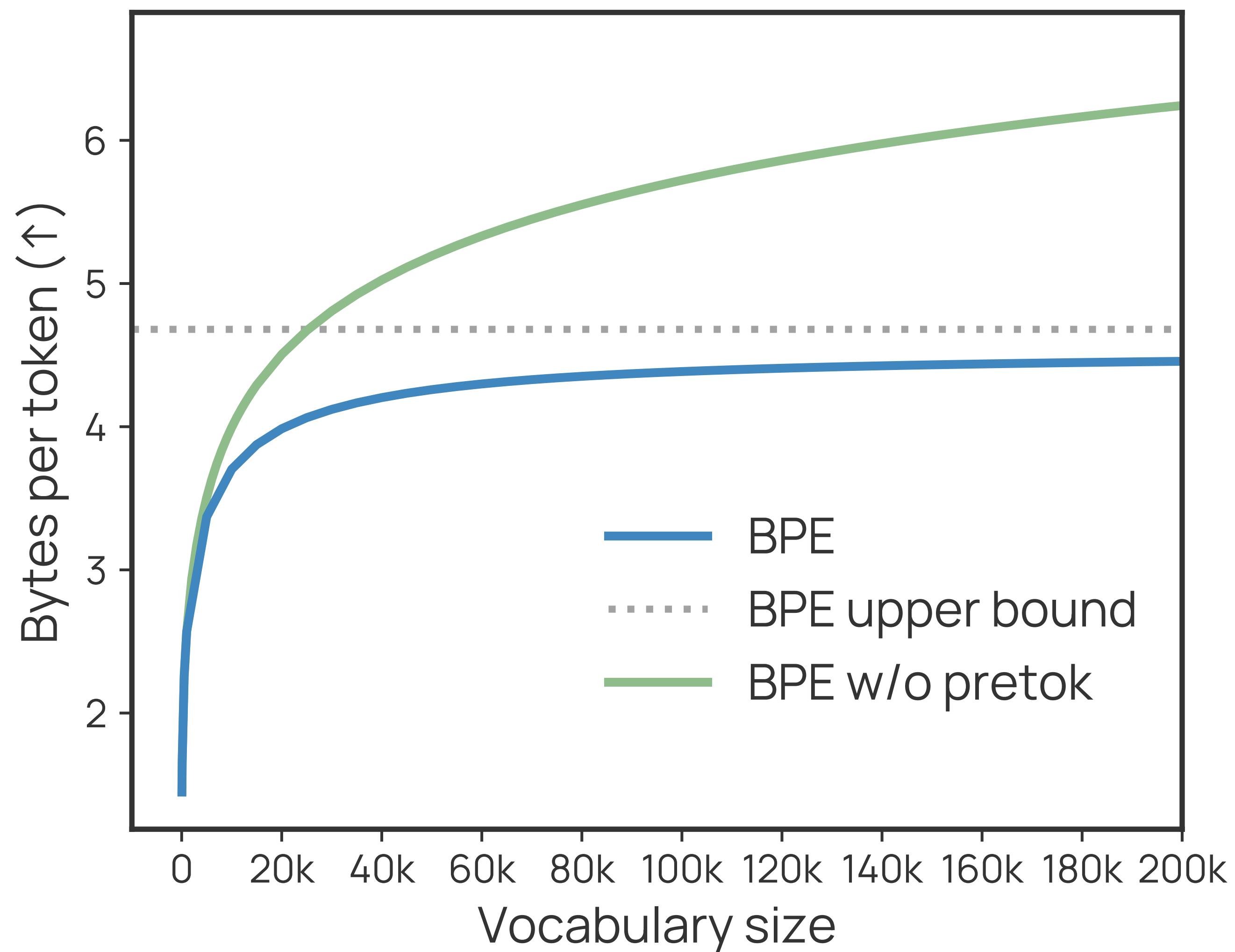
SuperBPE encodes text more efficiently



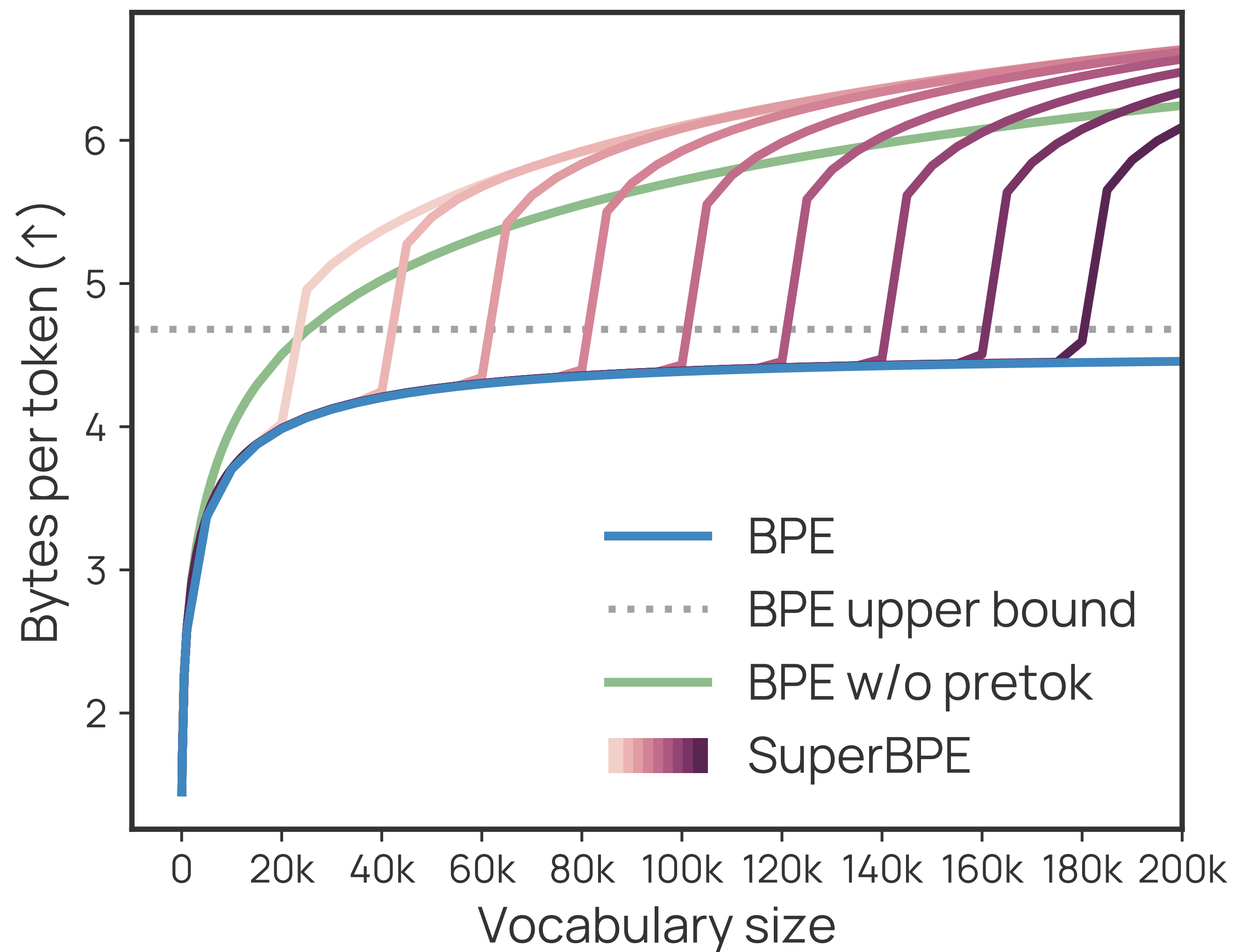
SuperBPE encodes text more efficiently



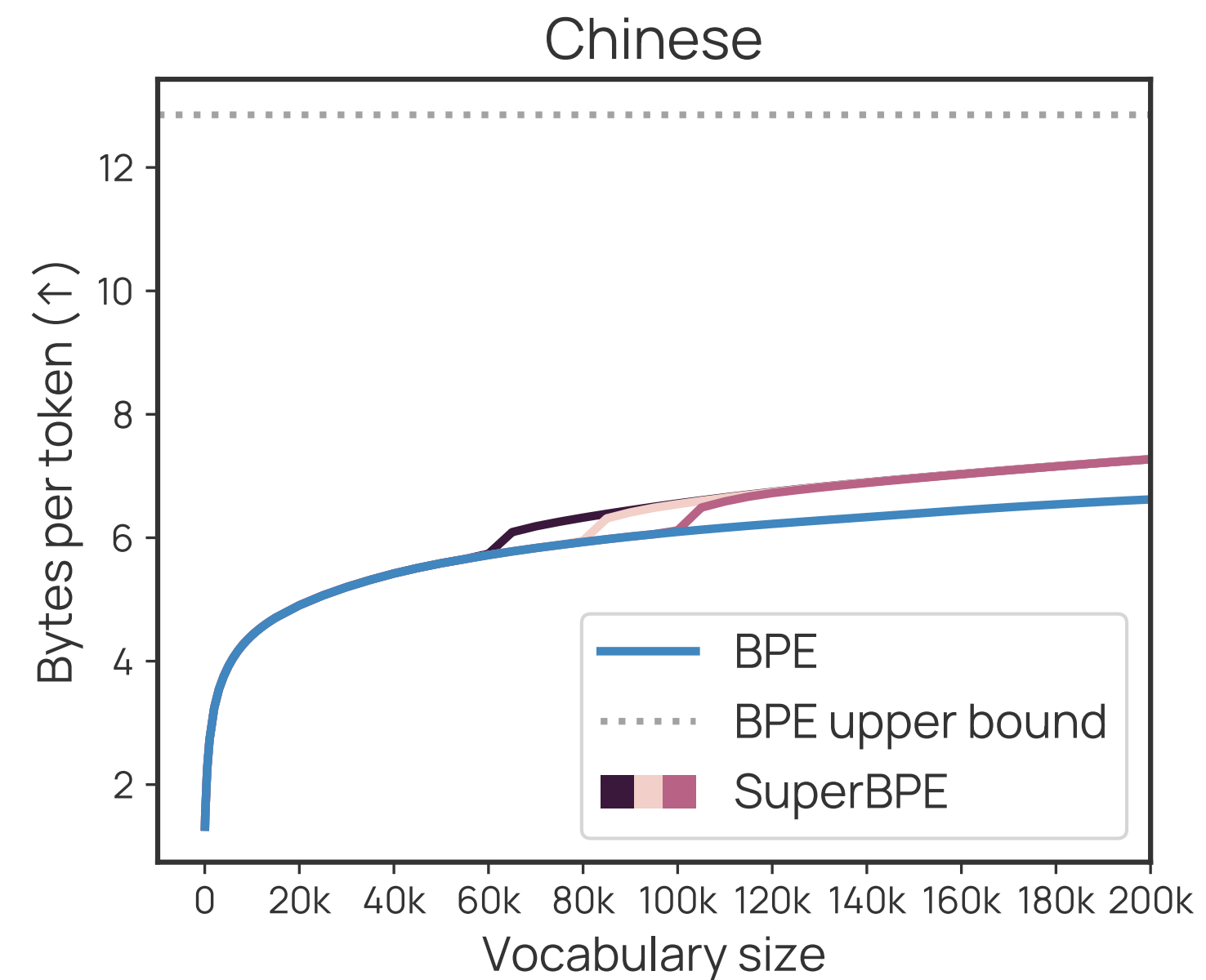
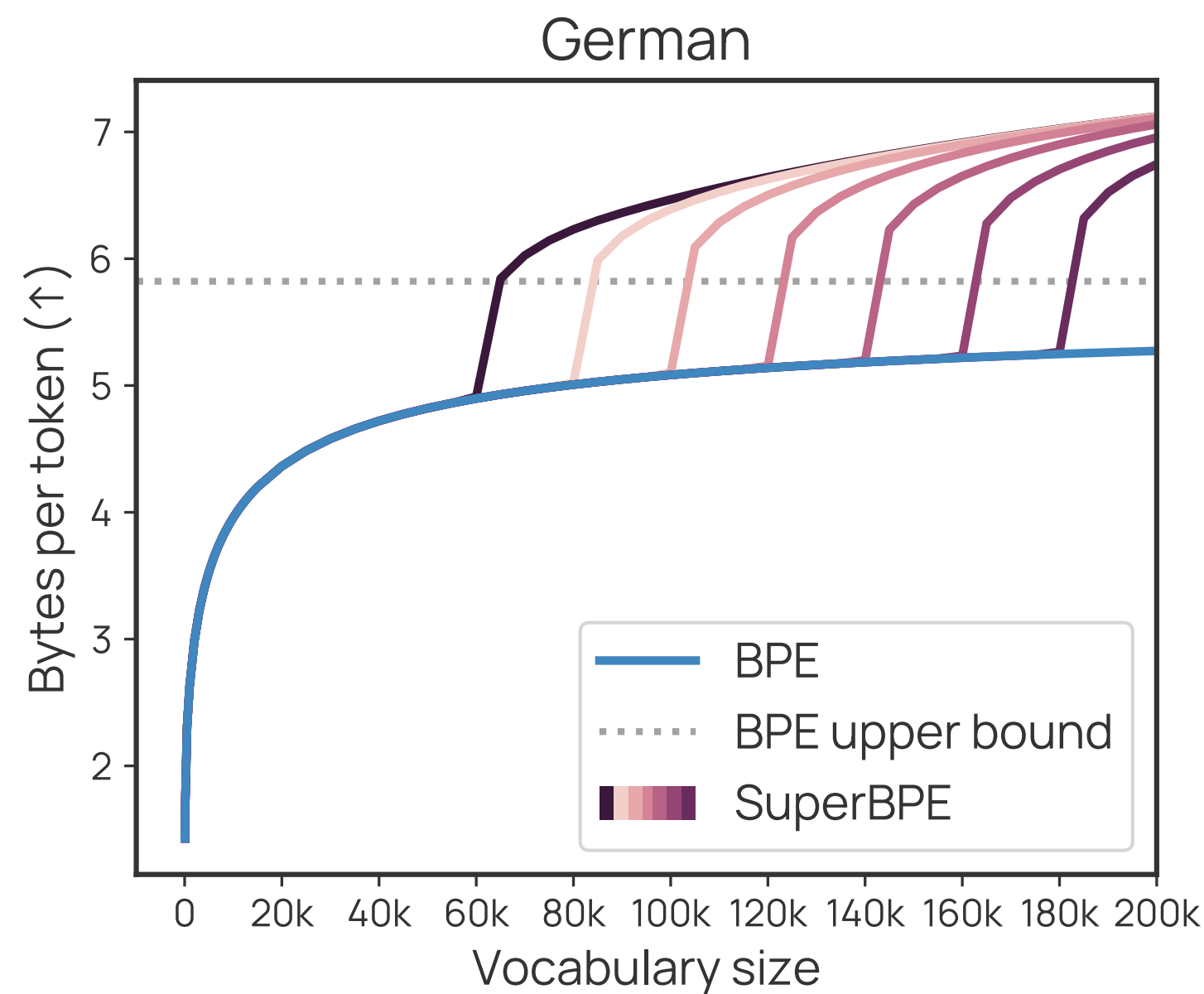
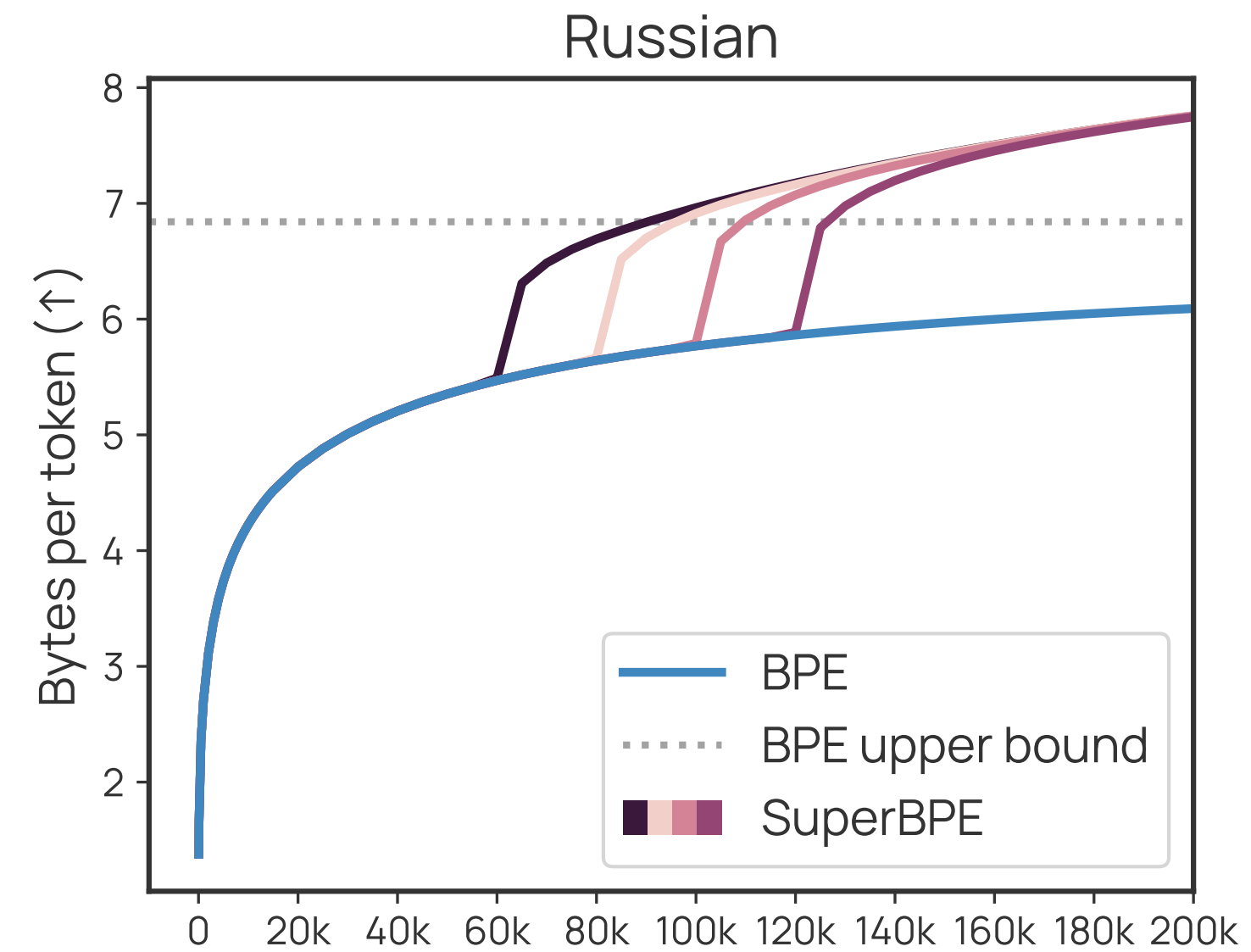
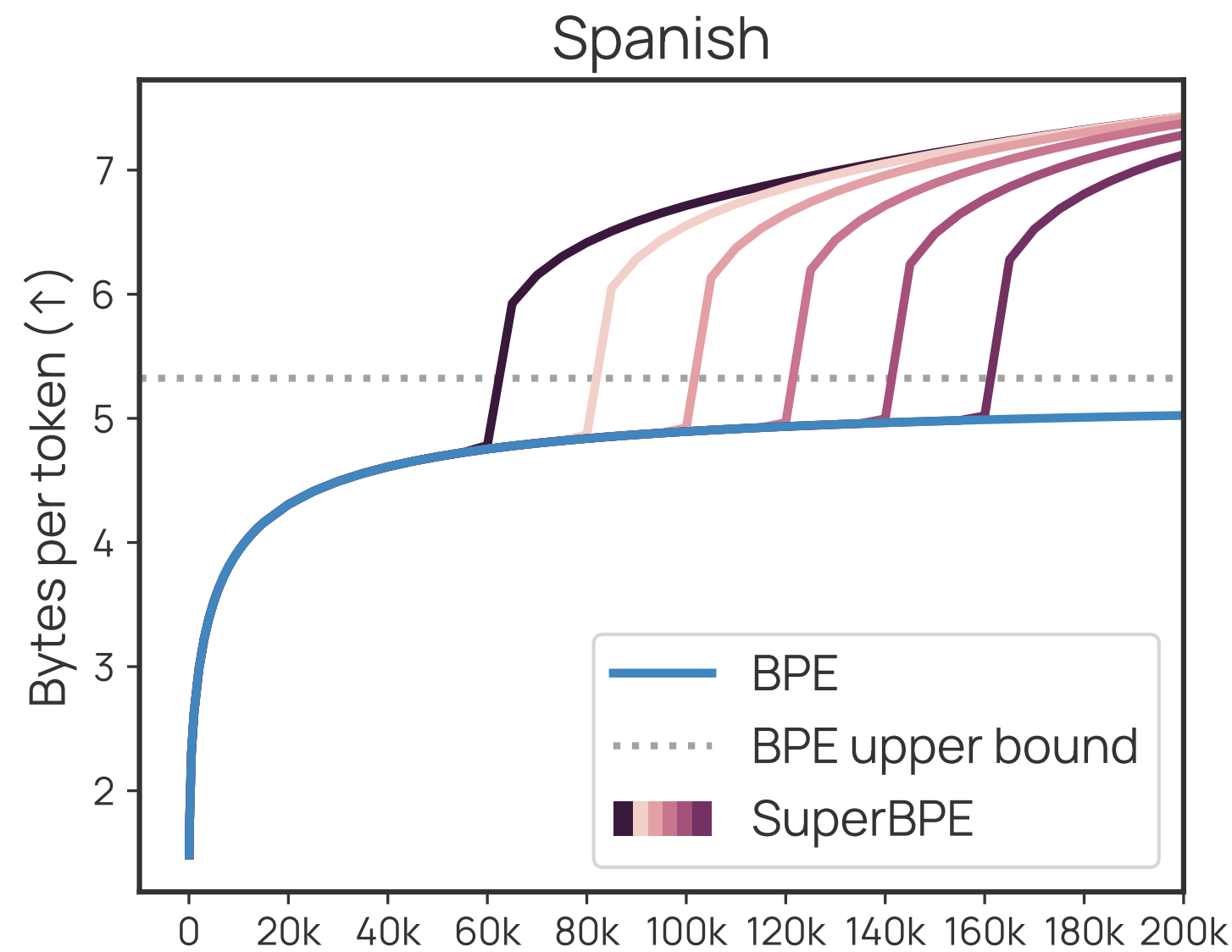
SuperBPE encodes text more efficiently



SuperBPE encodes text more efficiently



Efficiency scaling for non-English languages



Model training: what to control for?

There are many factors that can influence model performance!

- Model size
- Training compute
- Inference compute
- Amount of text seen (during training)
- Context length

How can we design a fair comparison?

It's impossible to hold everything constant!

Training compute seems like the most important thing to control for

Comparing LM loss across tokenizers

BPE

7 4 3 0 2 3 2 6 0 2 10 0 4

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

total loss = 43

avg loss = 3.3

total bytes = 40

bits per byte = 1.5

SuperBPE

14 5 2 6 2 10 4

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

total loss = 43

avg loss = 6.1

total bytes = 40

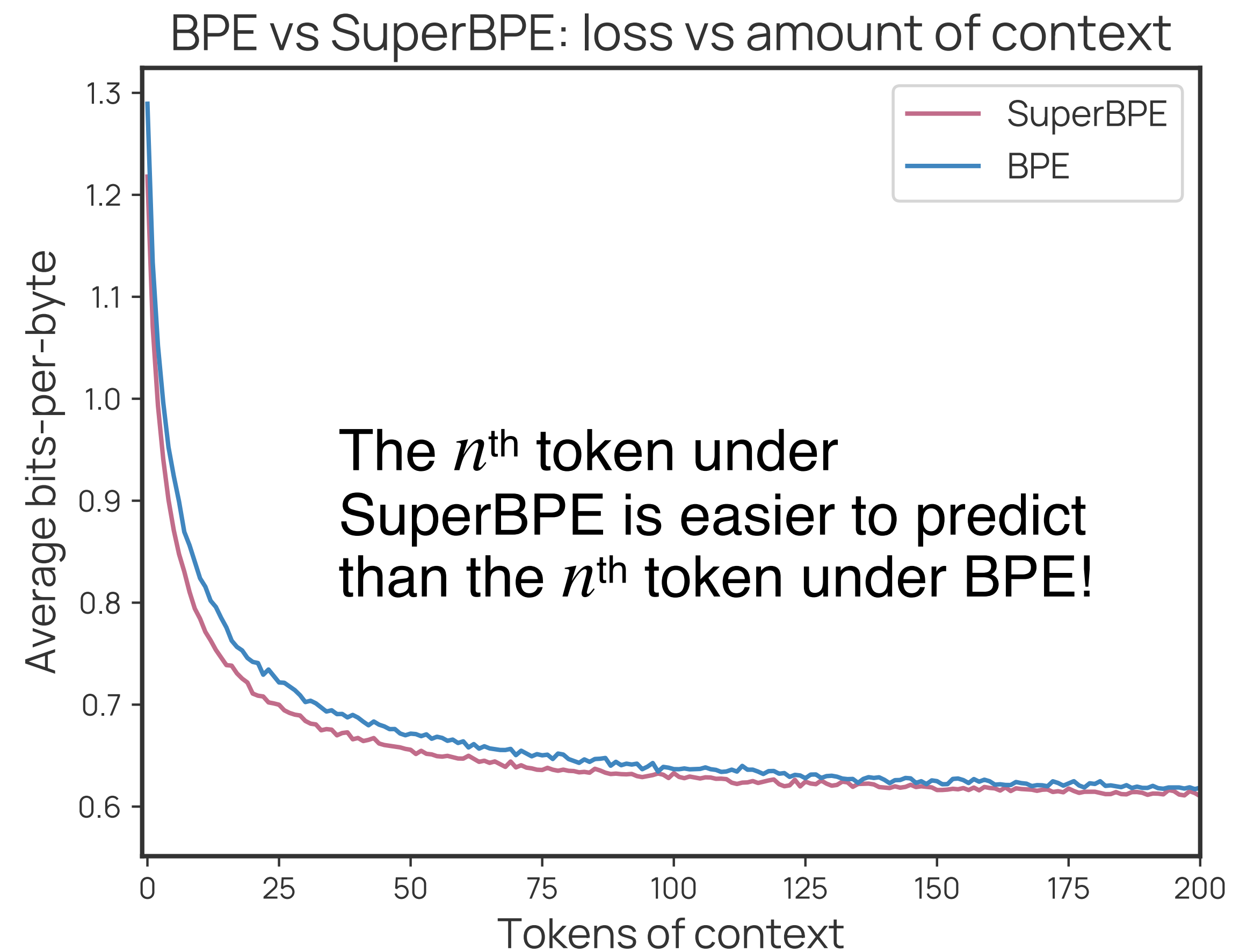
bits per byte = 1.5

How to control for context length

It's easier to predict text when more context is available

Our tokenizer sees more text per token

To avoid unfair advantages, we reduce the context length of our model in *tokens* to match the effective context length in *bytes*

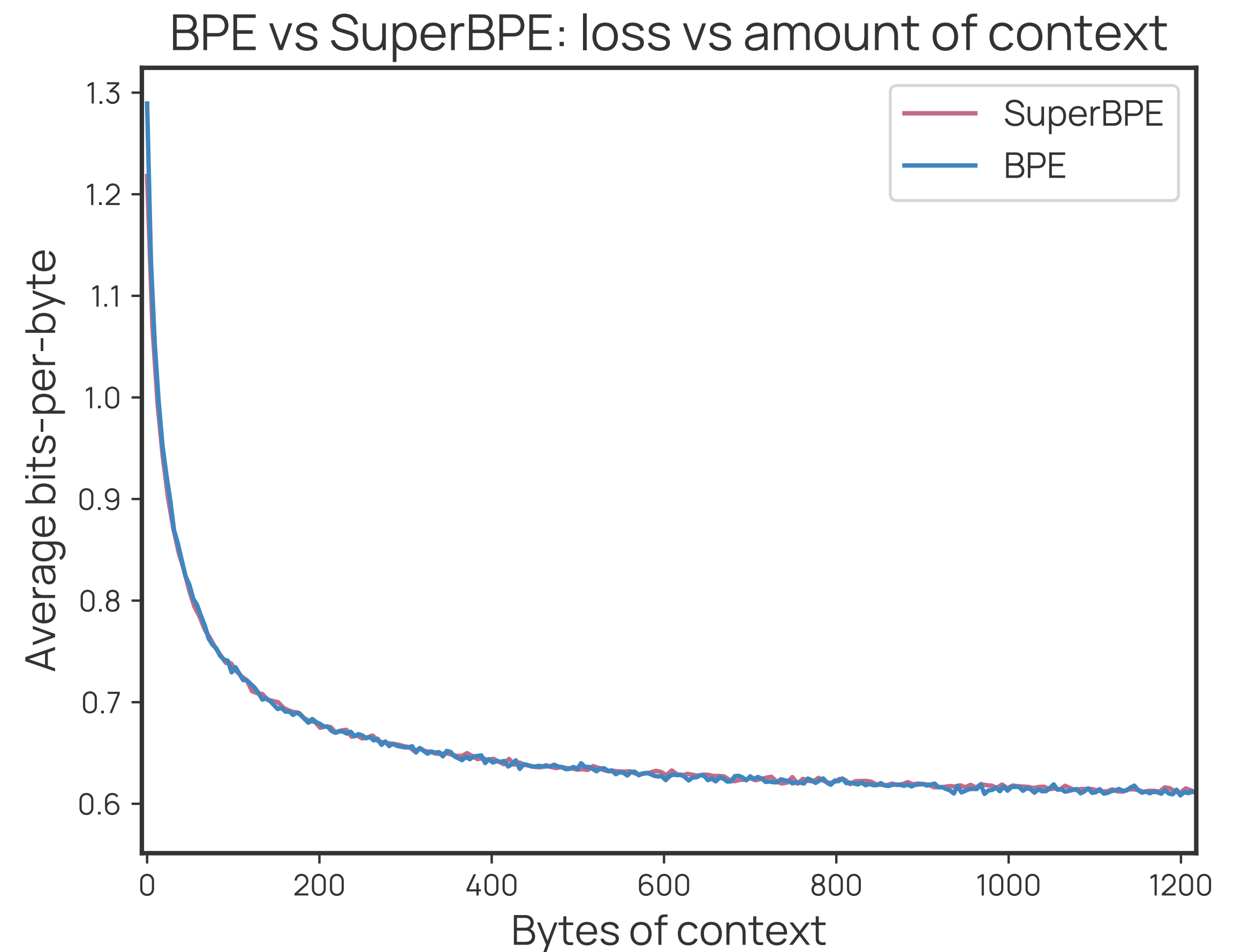


How to control for context length

It's easier to predict text when more context is available

Our tokenizer sees more text per token

To avoid unfair advantages, we reduce the context length of our model in *tokens* to match the effective context length in *bytes*



Settings we consider

Compared to an **8B BPE baseline** (Olmo2 @ 330B tokens)...

Option 1: SuperBPE 8B

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

“I don’t want the model to be bigger.”

— somebody

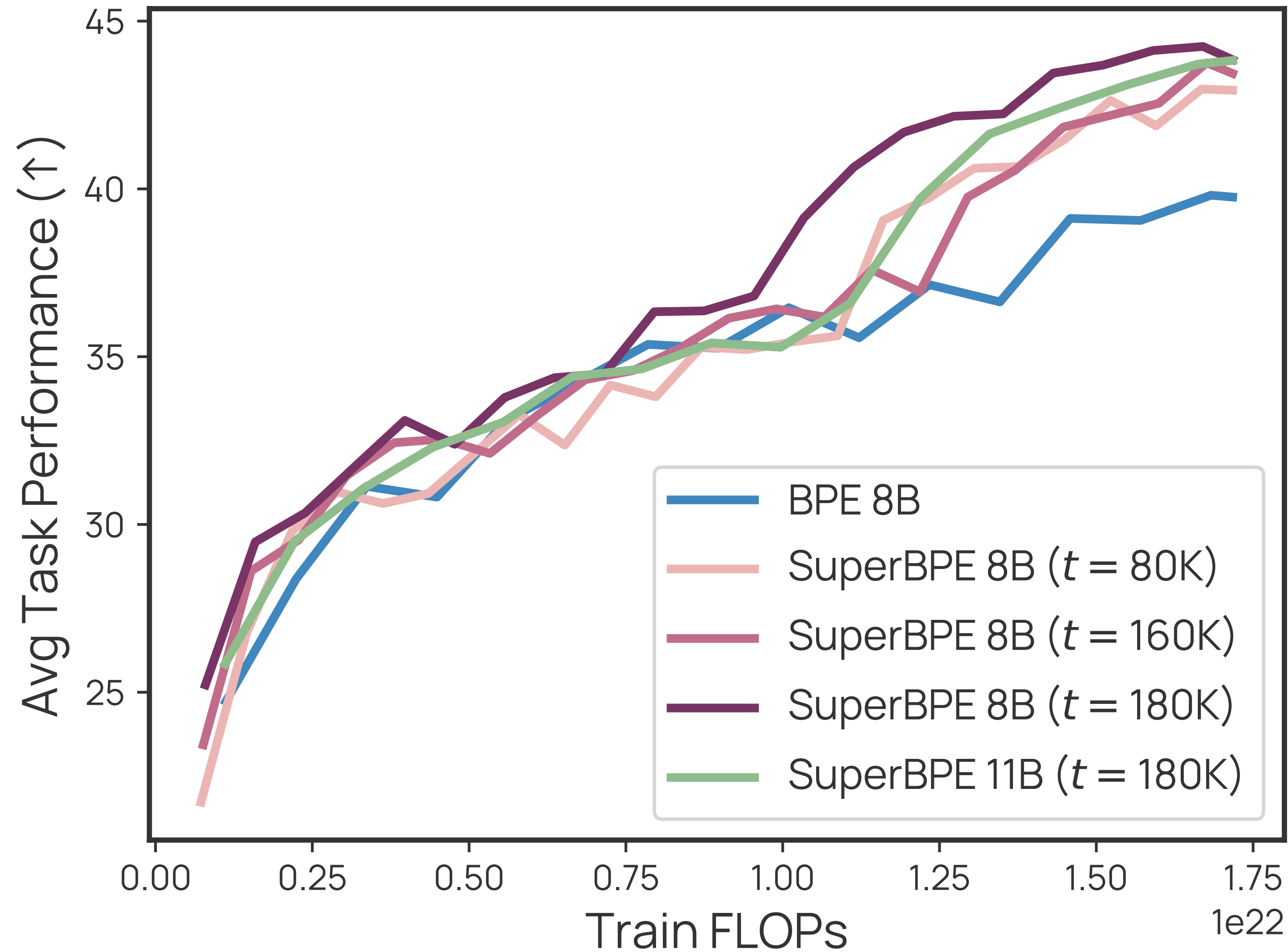
Option 2: SuperBPE 11B

- ✗ model size (39% bigger)
- ✓ training compute
- ✓ inference compute
- ✓ amount of text seen

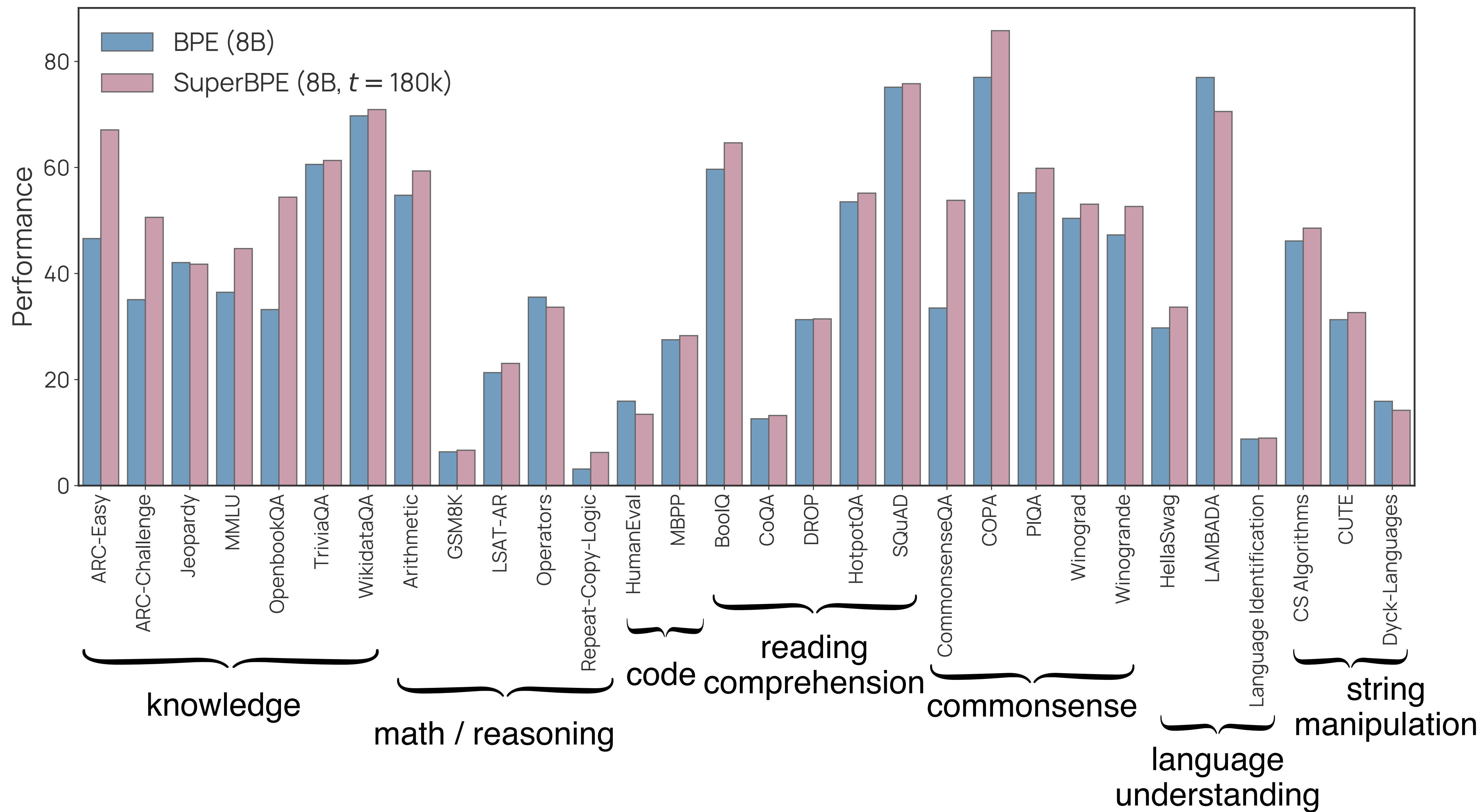
“I don’t have more training data.”

— somebody else

Task average during pretraining



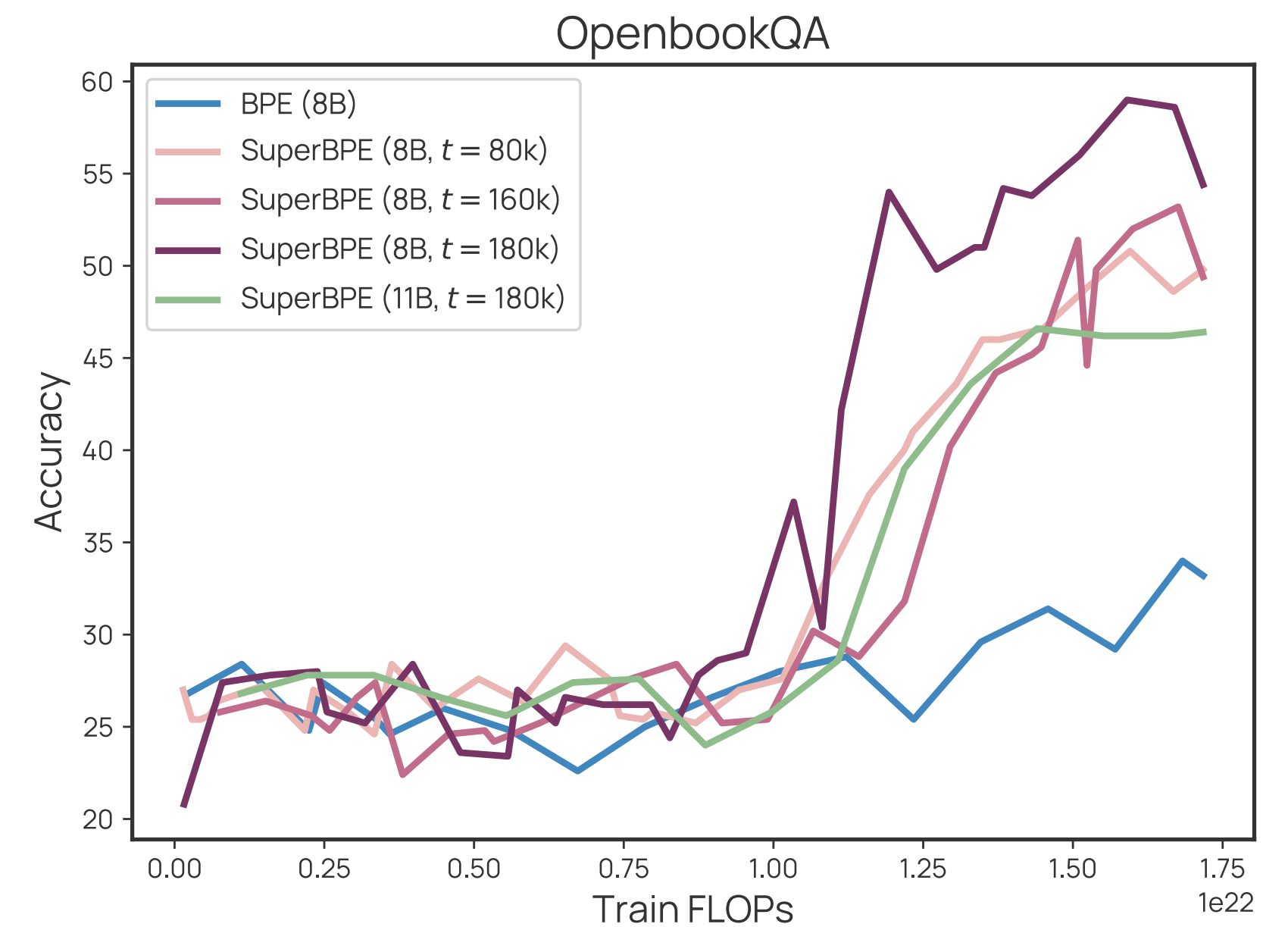
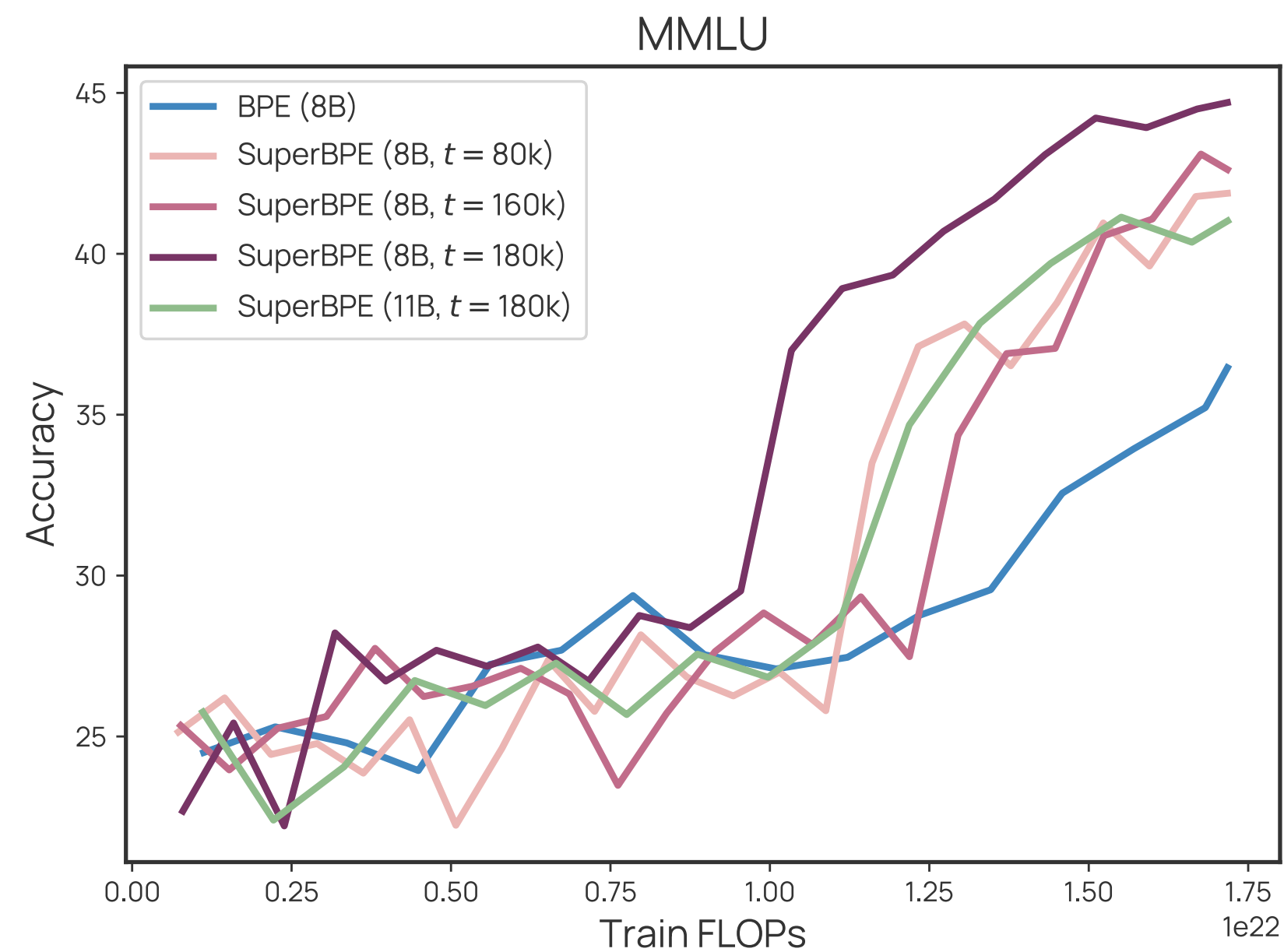
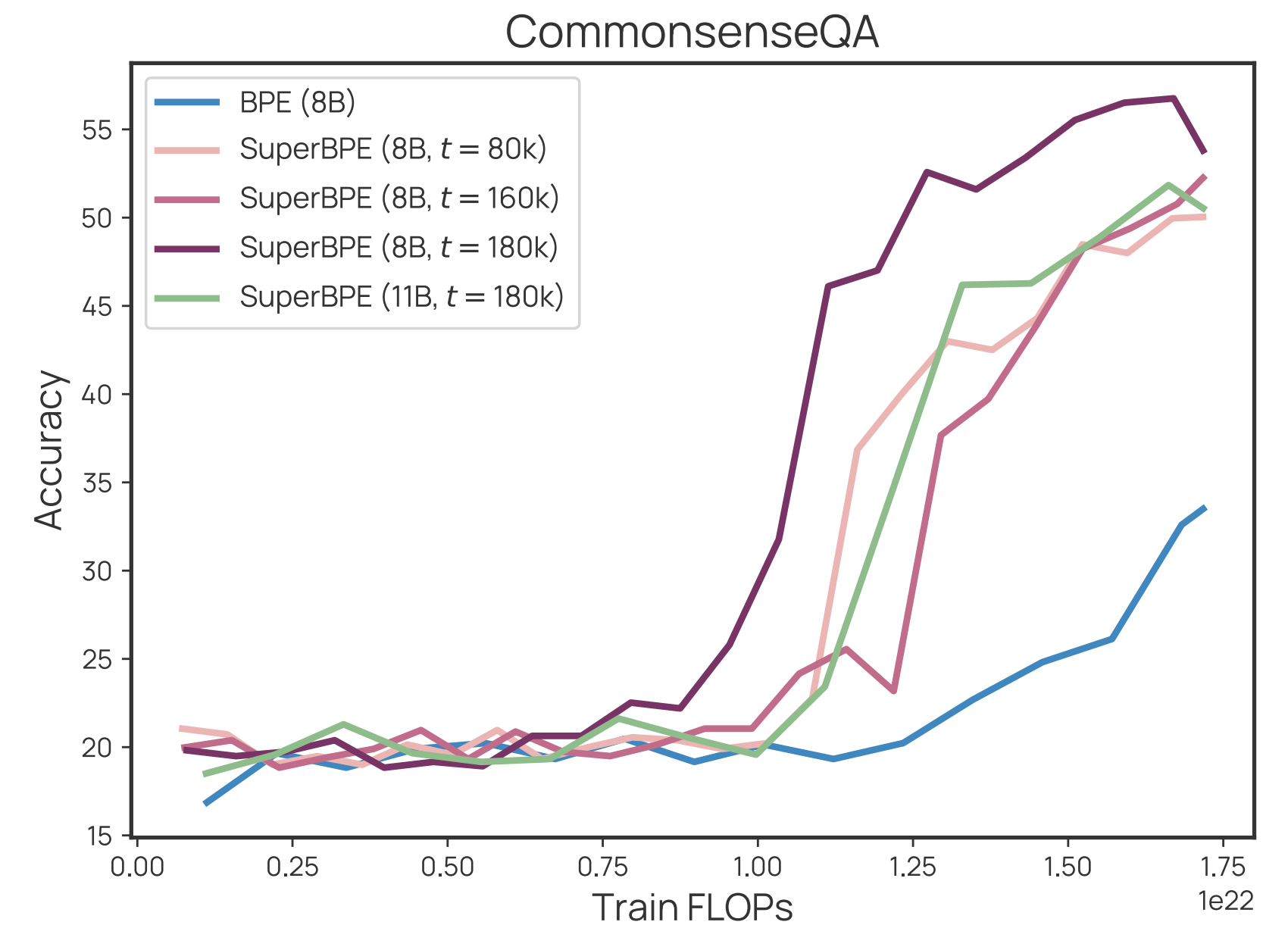
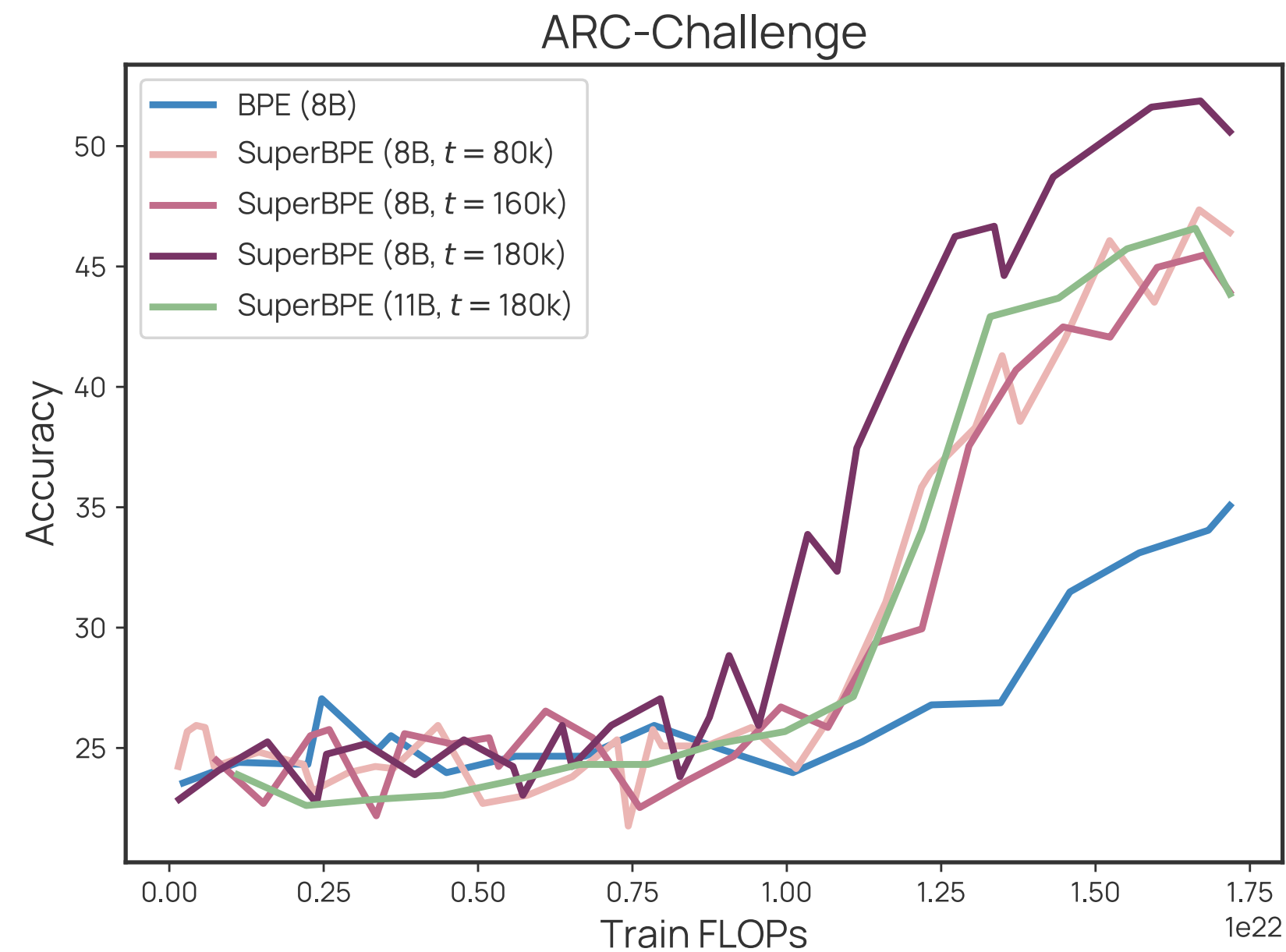
All SuperBPE models substantially outperform the baseline!



MC tasks

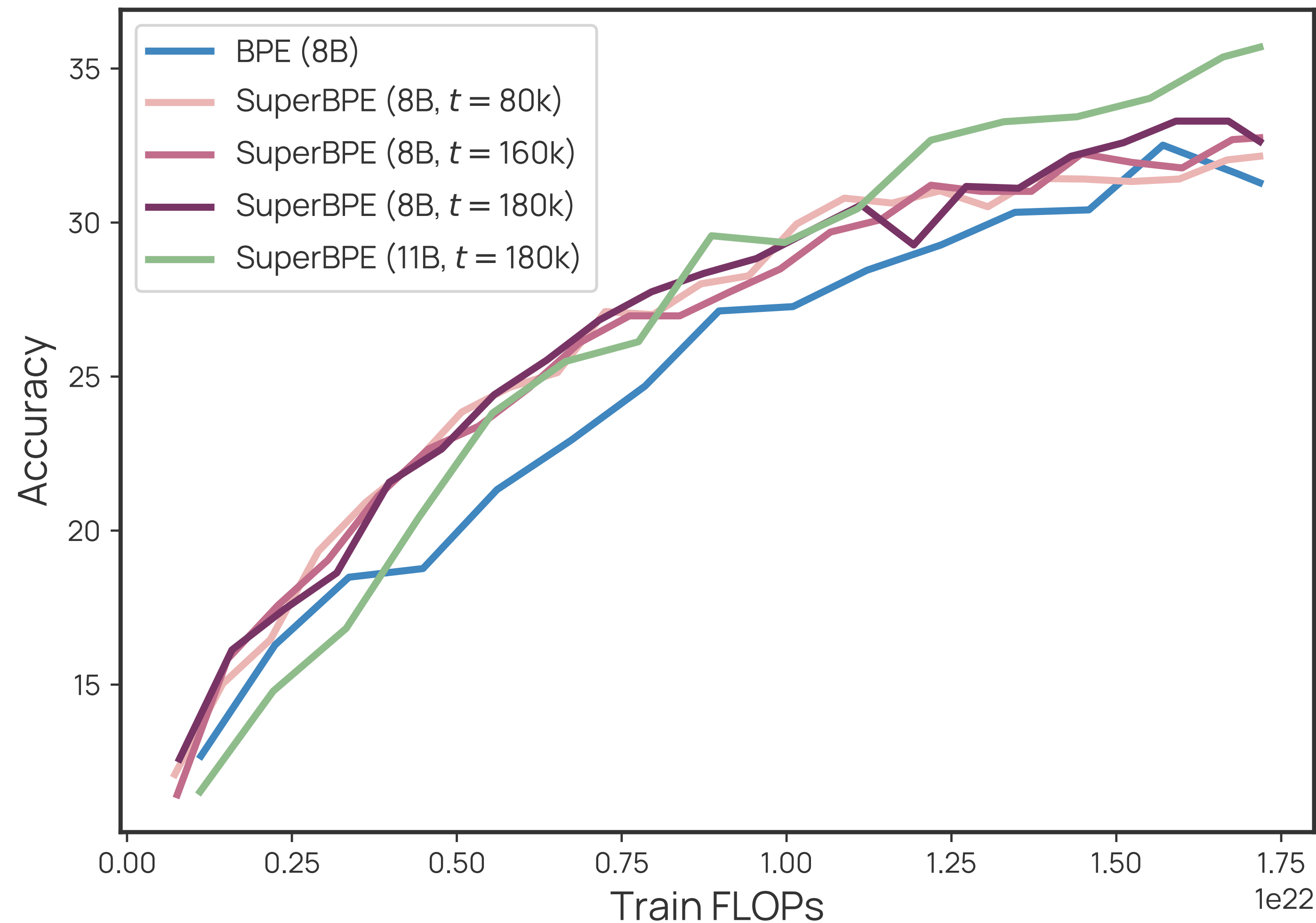
SuperBPE achieves large improvements in MC

All models begin to achieve better-than-random performance at a particular moment



String manipulation

CUTE



Is there a " c " in " also "?

*Delete every instance of "
t " in " data ".*

*Swap " k " and " e " in "
make ".*

What do tokens look like?

A lot of multi-word expressions (but not all are!)

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

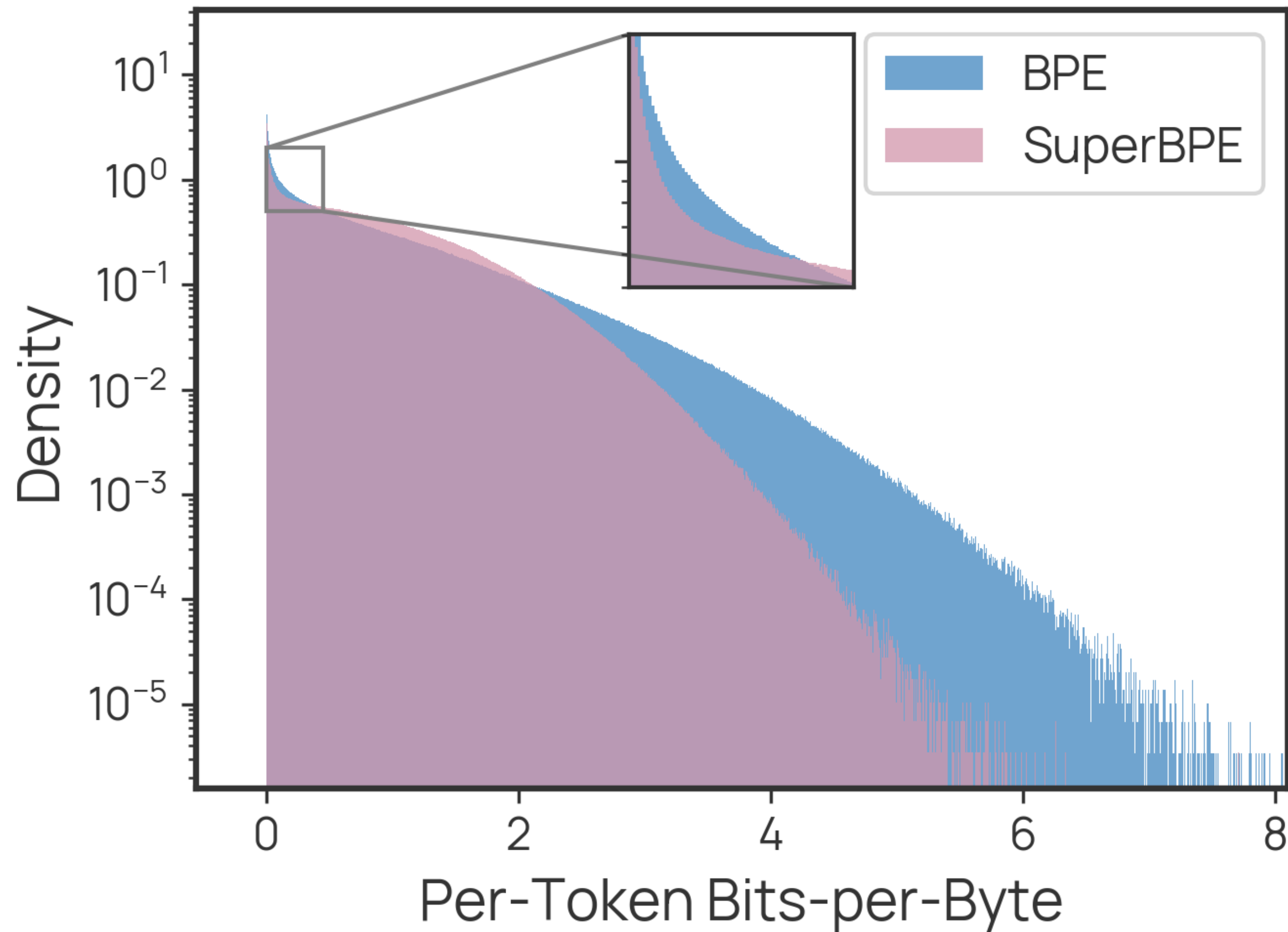
Language modeling performance

If you measure the loss of our model, you'll find that we're actually very slightly behind in bits-per-byte (0.0017)

	BPE	SuperBPE
bits-per-byte	0.7465	0.7482

That's interesting because we're winning on evals!

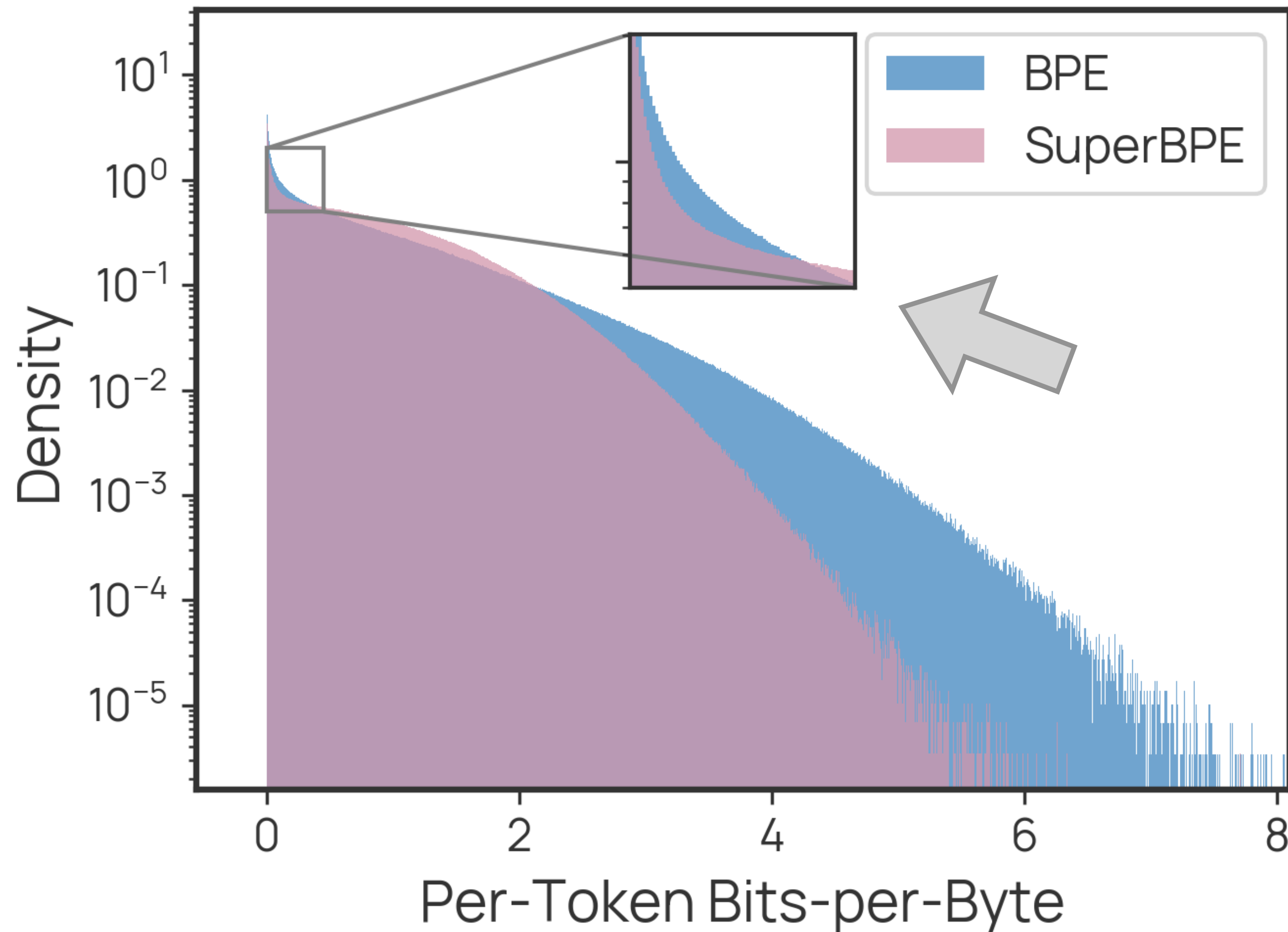
BPB Distribution



Mean BPB is very close
(SuperBPE behind by 0.0017)

But SuperBPE distributes loss
more uniformly over tokens

BPB Distribution

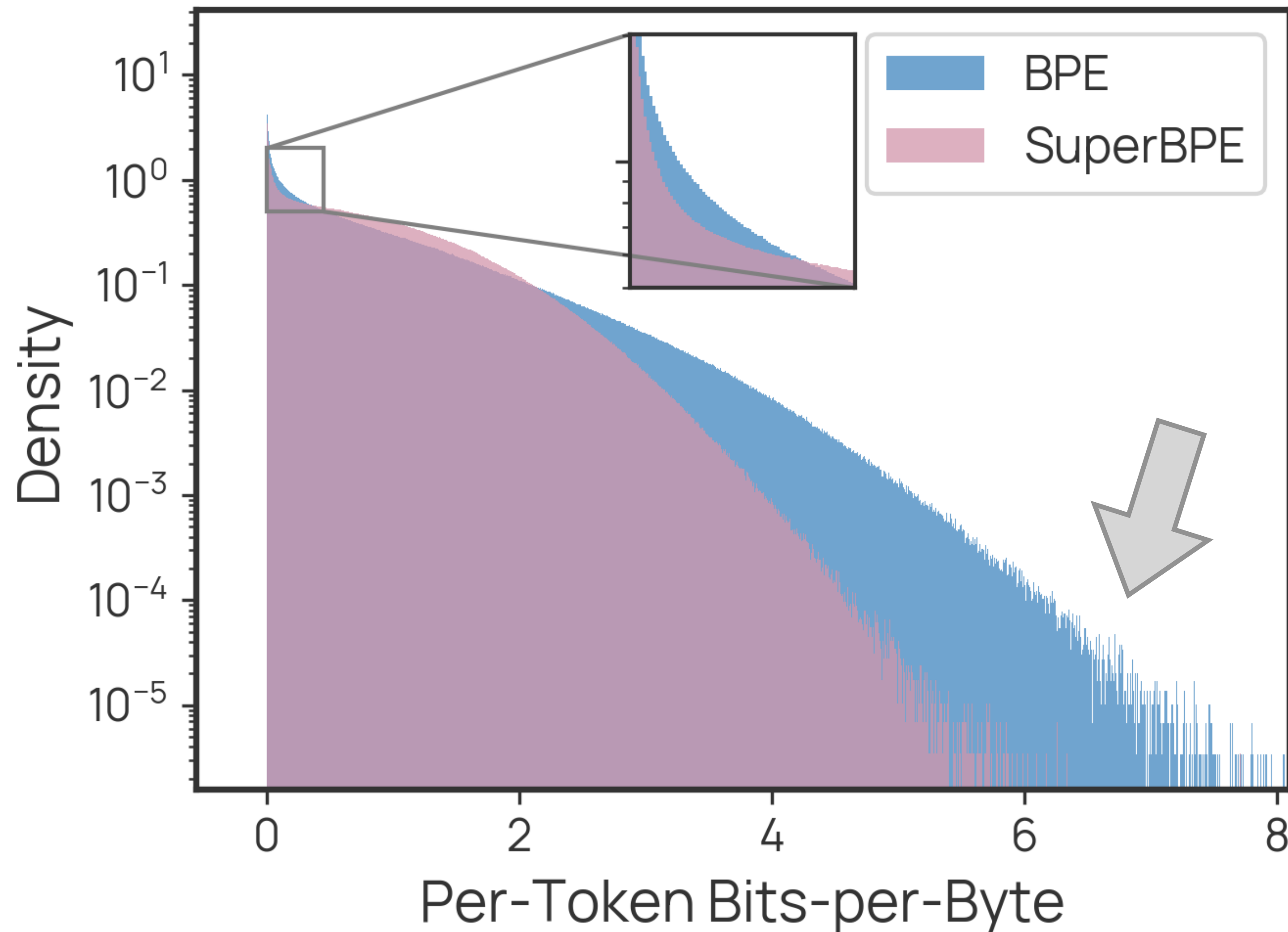


SuperBPE has fewer very-low-loss tokens:

The token `_to` appears 10x more under BPE than SuperBPE, with average BPB 0.285 (for BPE) < 0.476 (SuperBPE)

By just excluding `_the`, `_to`, `_of` (3 most common words), BPB ranking reverses

BPB Distribution



SuperBPE has fewer very-high-loss tokens:

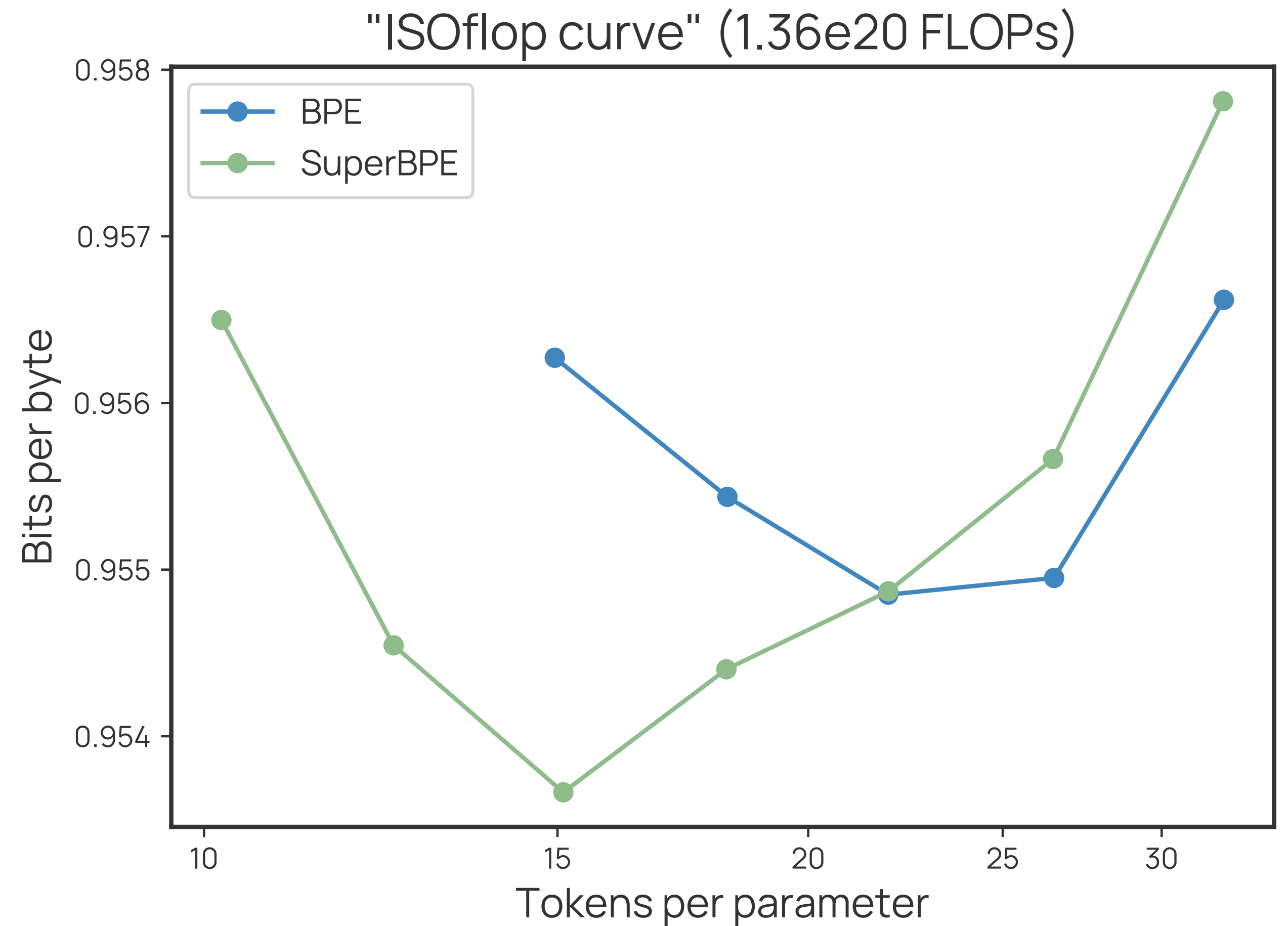
May explain why we win on evals (evals are hard)

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

Scaling part 1

Prior work on scaling finds that ~20-25 tokens per parameter (T/P) achieves the lowest loss when keeping training compute fixed

But our tokens are harder to predict, so maybe the optimal T/P ratio will be different?

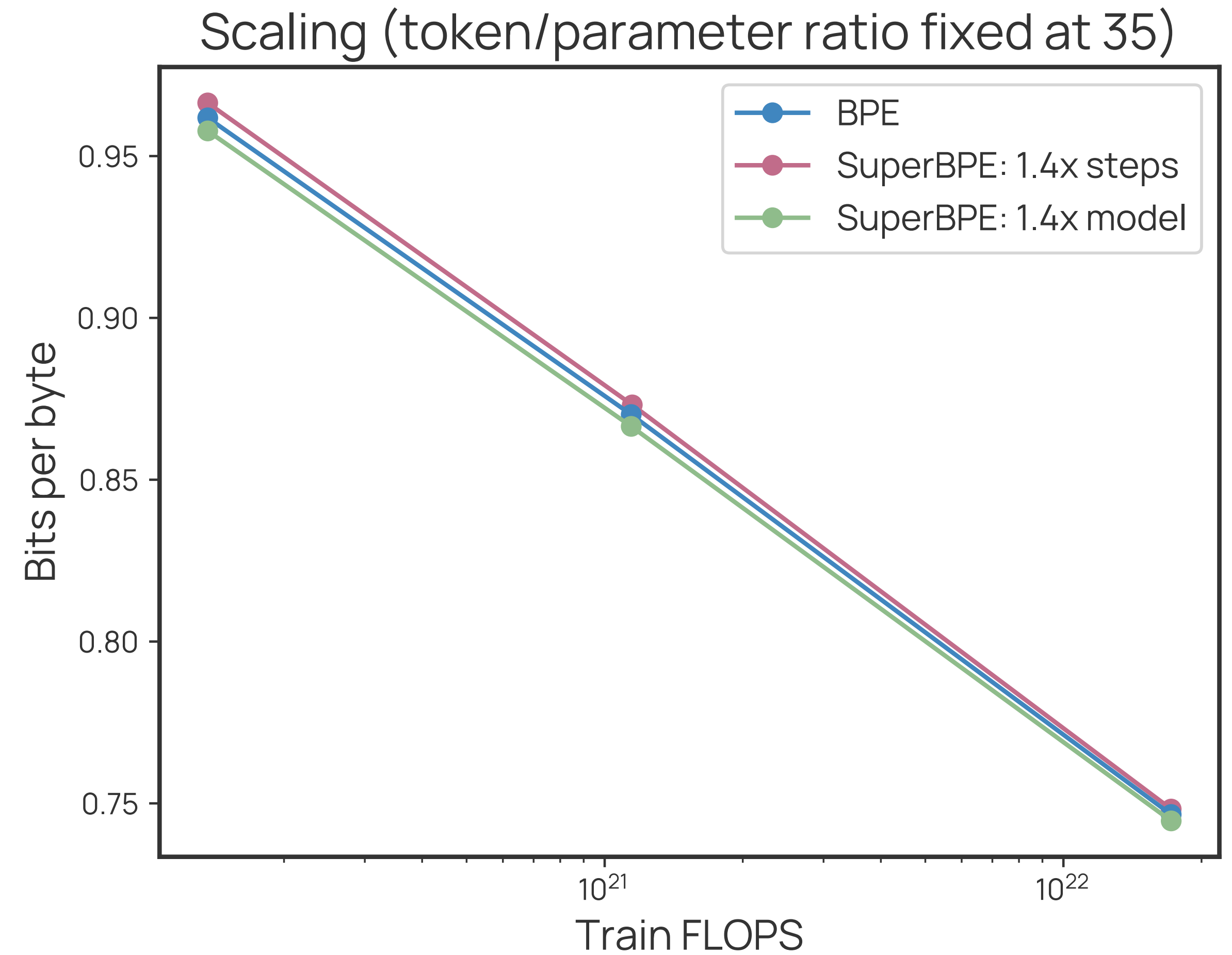


Training Compute-Optimal LLMs (Hoffman et al., 2022)

Scaling Laws for Neural LMs (Kaplan et al., 2020)

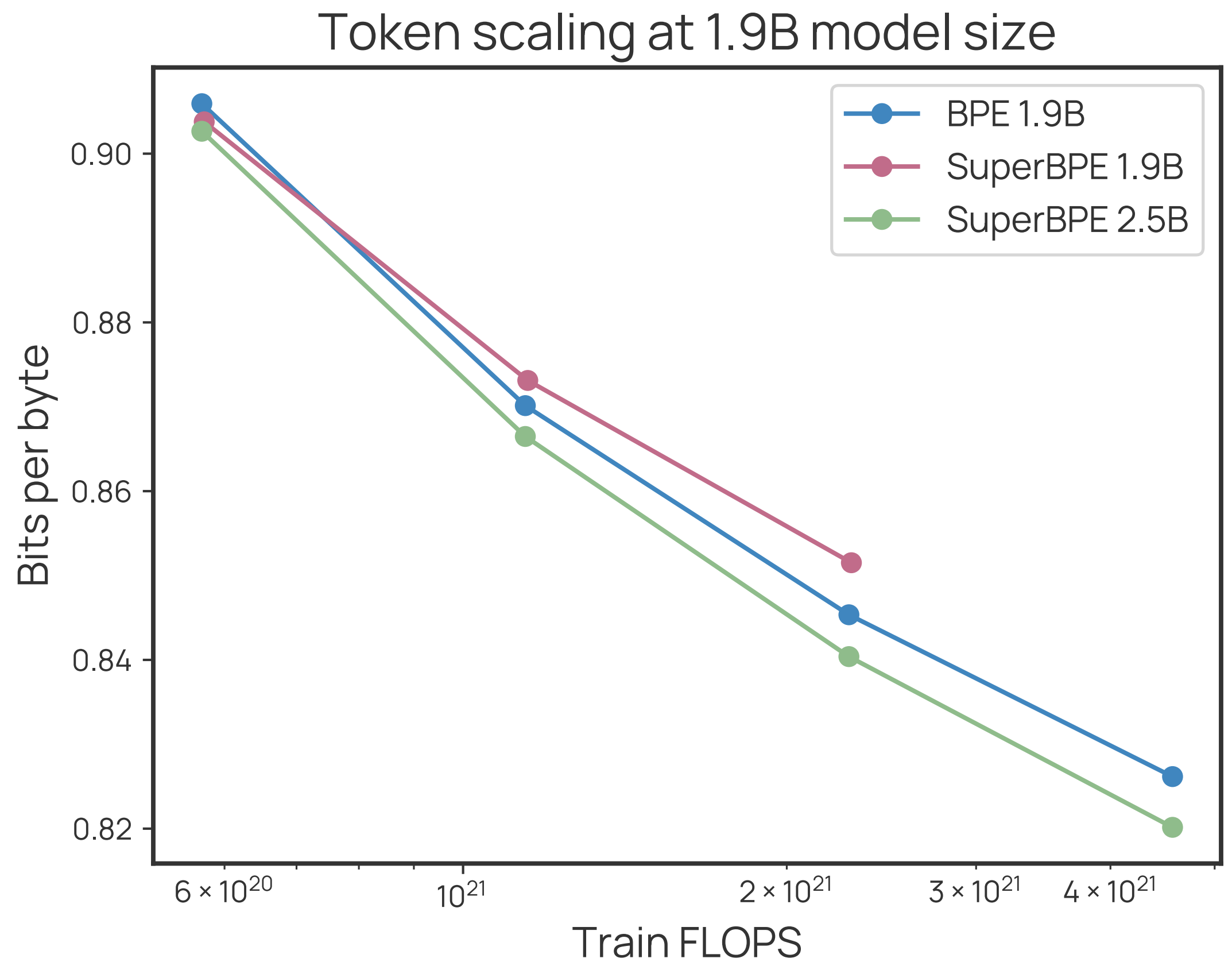
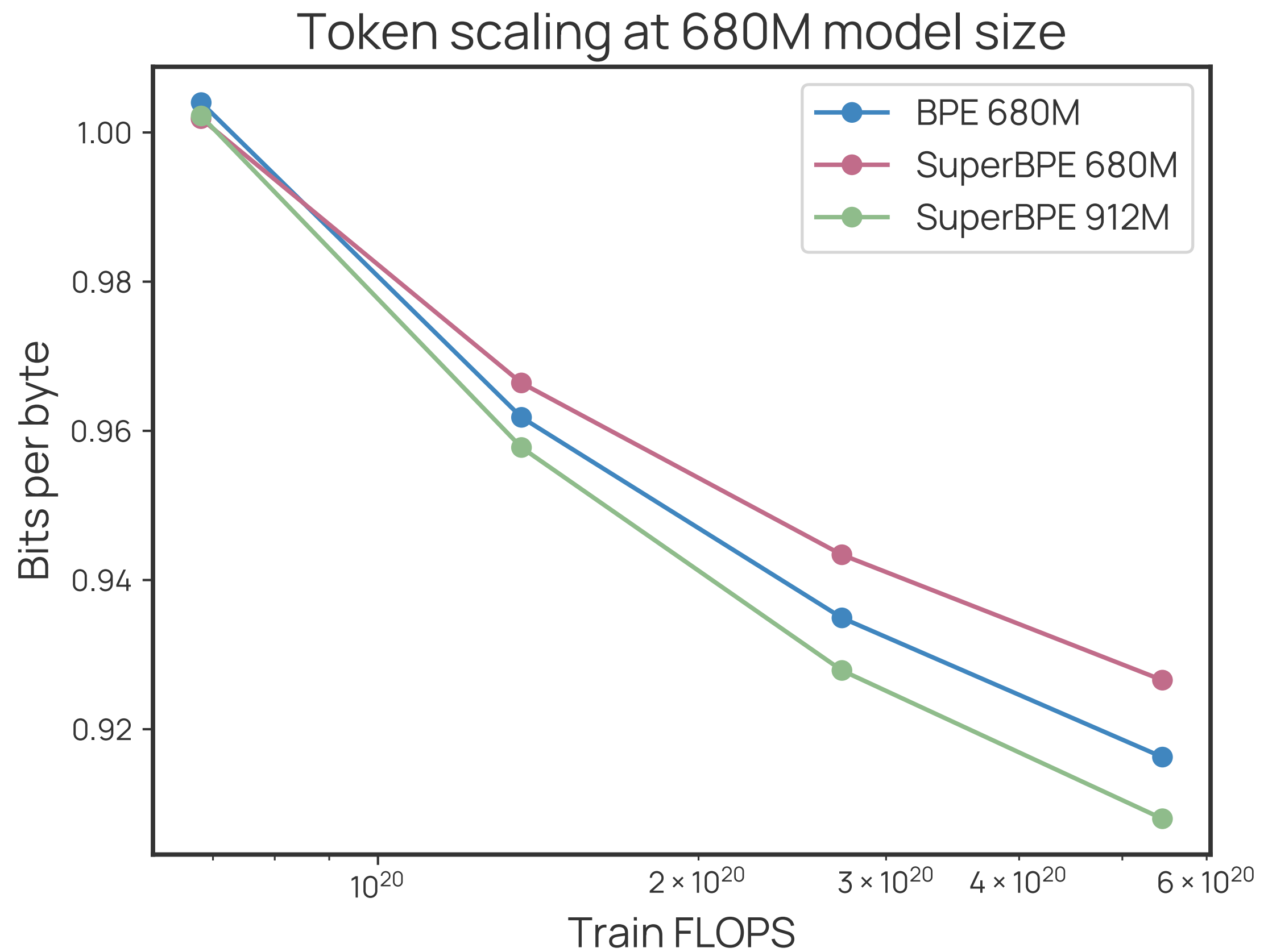
Scaling part 2

Ok, what happens if we fix the token/parameter ratio and scale up?



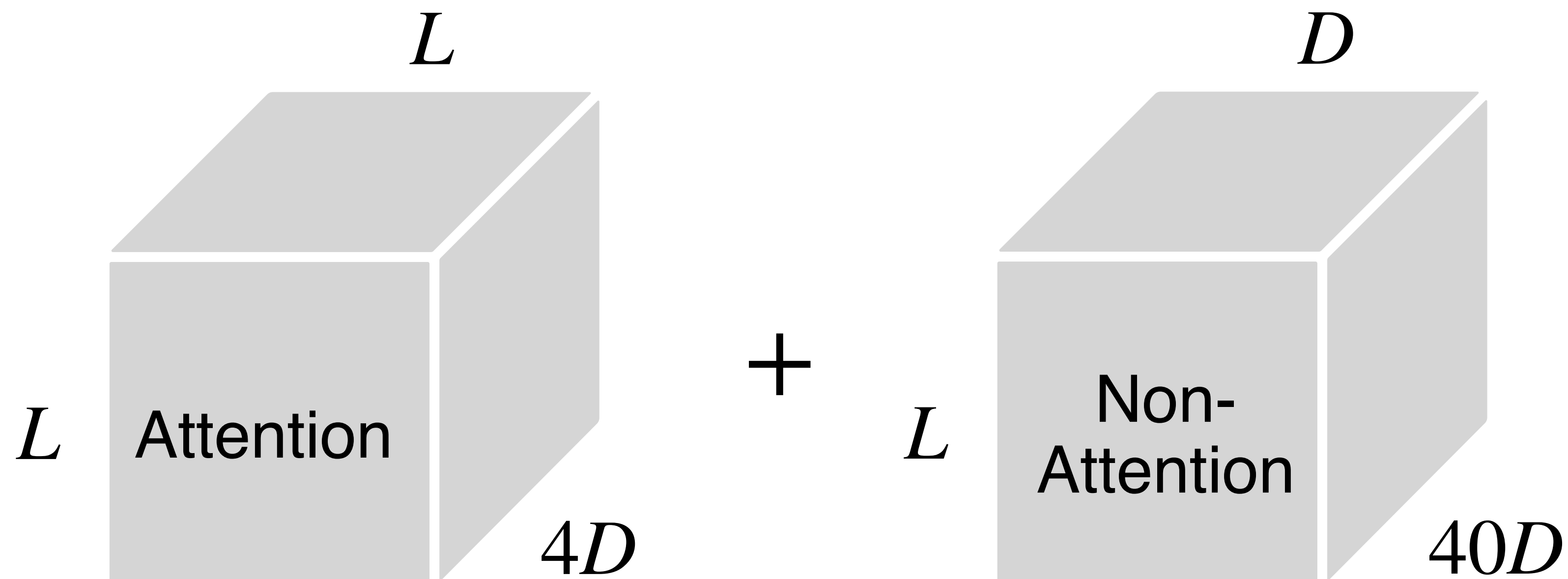
Scaling part 3

But people are mostly interested in overtrained models!



Scaling the context length

How do the flops scale? Let D be the model hidden dimension and L be the max sequence length. At every layer, we require this many FLOPs:



assuming MLP expansion factor of 8 (typical)

Takeaways

SuperBPE extends subword BPE to include *superwords*, or tokens containing (parts of) multiple words!

SuperBPE encodes the same text using dramatically fewer tokens than BPE

Given fixed train compute, we can train on more text *or* train a bigger model, both of which consistently bring gains in downstream evaluations

SuperBPE models distribute loss more evenly over tokens

