Tokenization: How do language models see text?

Jan 27, 2025

CSE 447/517: NLP

Guest lecture from Alisa Liu

Inspiration taken from lectures of Yejin Choi, Andrej Karpathy, Sachin Kumar, Oreva Ahia

Tokenization :(

Tokenization is at the heart of much weirdness of LLMs. Do not brush it off.

- Why can't LLM spell words? **Tokenization**.
- Why can't LLM do super simple string processing tasks like reversing a string? **Tokenization**.
- Why is LLM worse at non-English languages (e.g. Japanese)? **Tokenization**.
- Why is LLM bad at simple arithmetic? **Tokenization**.
- Why did GPT-2 have more than necessary trouble coding in Python? **Tokenization**.
- Why did my LLM abruptly halt when it sees the string "<|endoftext|>"? **Tokenization**.
- What is this weird warning I get about a "trailing whitespace"? **Tokenization**.
- Why the LLM break if I ask it about "SolidGoldMagikarp"? **Tokenization**.
- Why should I prefer to use YAML over JSON with LLMs? **Tokenization**.
- Why is LLM not actually end-to-end language modeling? **Tokenization**.
- What is the real root of suffering? **Tokenization**.

Let's build the GPT Tokenizer

Andrej Karpathy
609K subscribers
Subscribe

Outline

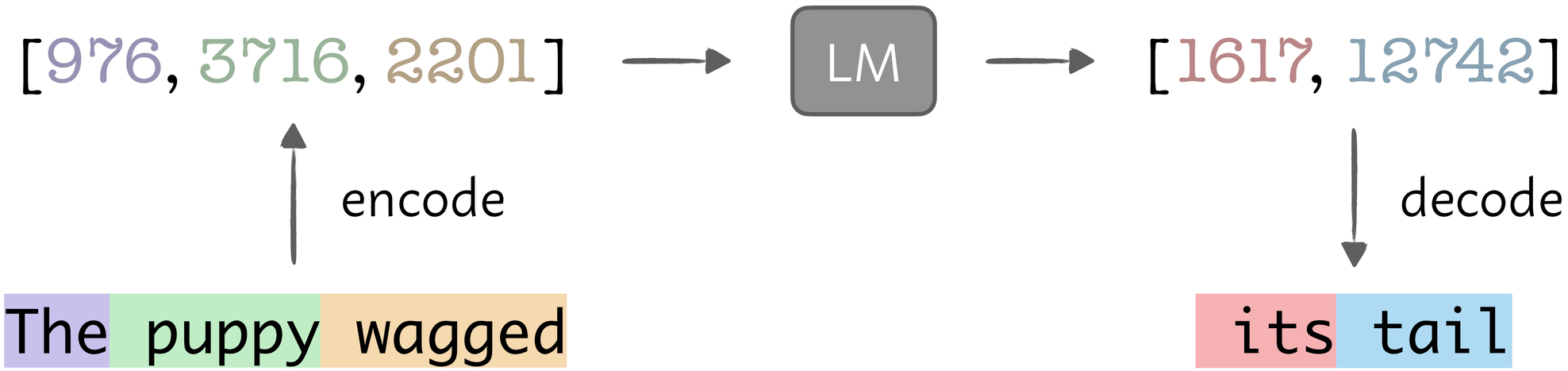
1. What is tokenization?
2. Word-level and character-level tokenizers
3. Subword-level tokenizers
4. BPE: Byte Pair Encoding
5. Variations on tokenization

What is tokenization?

Token = a unit with its own embedding representation

A **tokenizer** translates between text and a sequence of **tokens** that a language model (LM) learns over

The **vocabulary** V is the set of known tokens

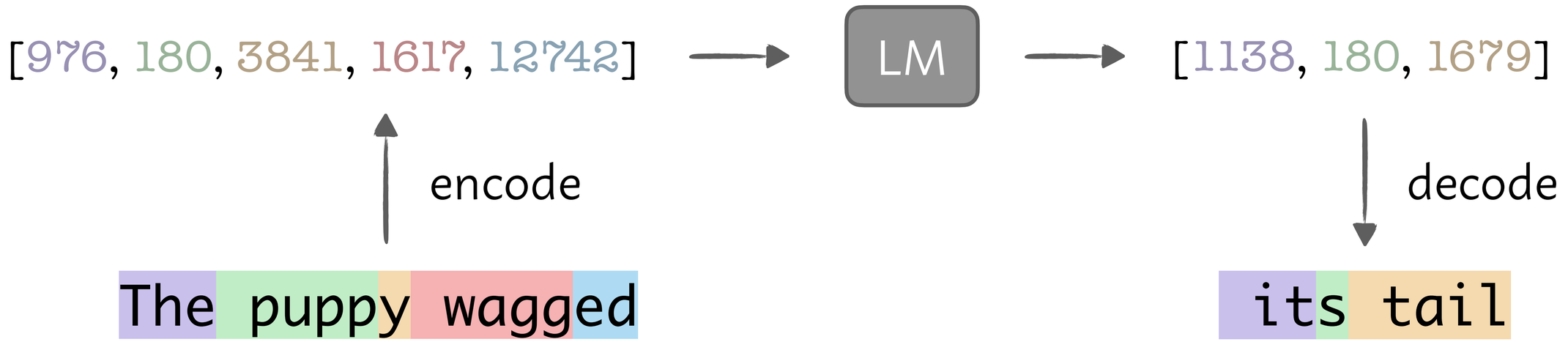


What is tokenization?

Token = a unit with its own embedding representation

A **tokenizer** translates between text and a sequence of **tokens** that a language model (LM) learns over

The **vocabulary** V is the set of known tokens

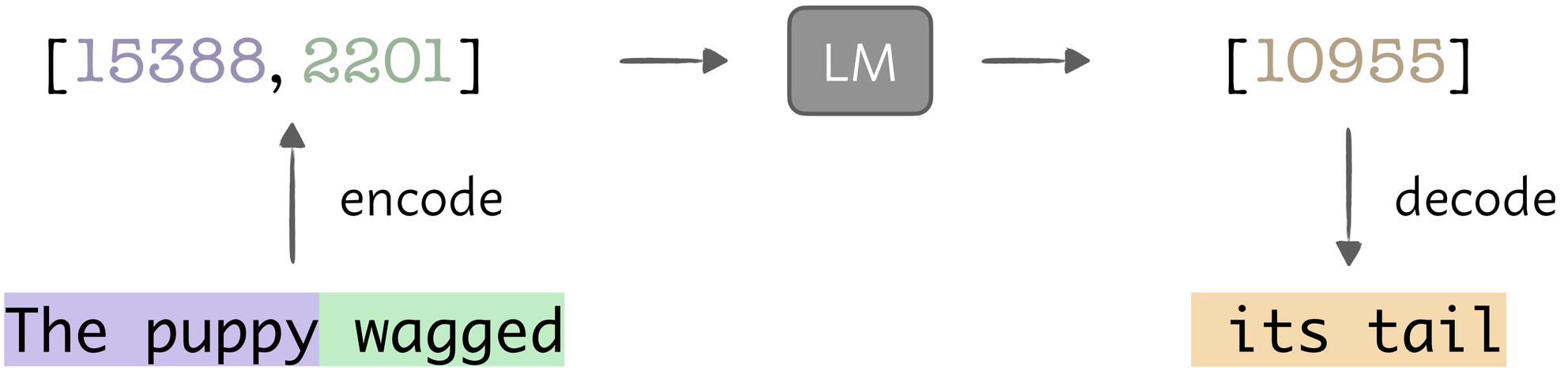


What is tokenization?

Token = a unit with its own embedding representation

A **tokenizer** translates between text and a sequence of **tokens** that a language model (LM) learns over

The **vocabulary** V is the set of known tokens

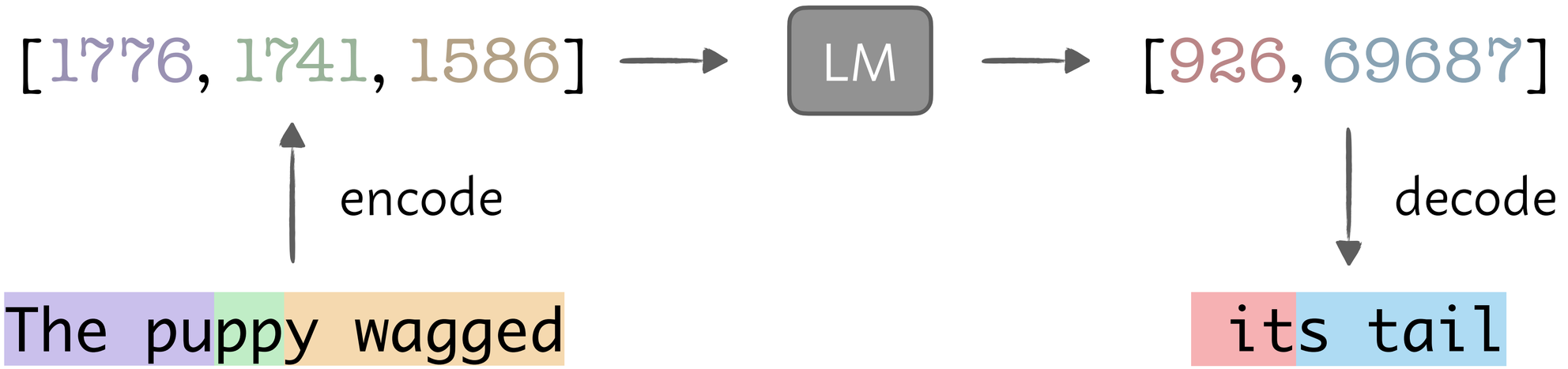


What is tokenization?

Token = a unit with its own embedding representation

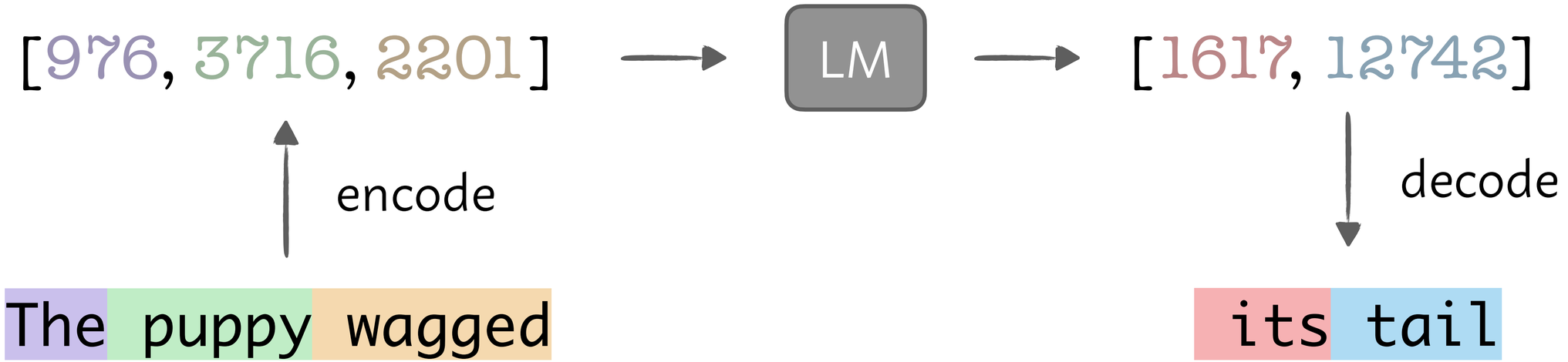
A **tokenizer** translates between text and a sequence of **tokens** that a language model (LM) learns over

The **vocabulary** V is the set of known tokens



Word-level tokenization

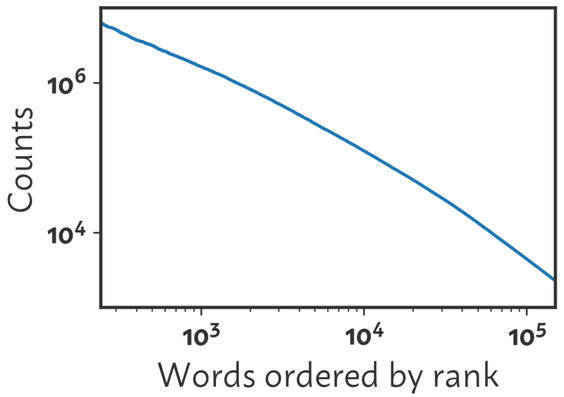
V = set of all words in the English language



Word-level tokenization

✅ Pros

Semantically meaningful language units



What does "breakfastish" mean?

"Breakfastish" is an informal and playful term that means "resembling or characteristic of breakfast." It's used to describe something that has qualities typically associated with breakfast, such as food items, timing, or atmosphere.

❌ Cons

$|V|$ can be quite large

- Webster's English dictionary has ~470,000 words!

Long tail of infrequent words

- **Zipf's law**: word freq. is inversely prop. to rank

Language is changing all the time

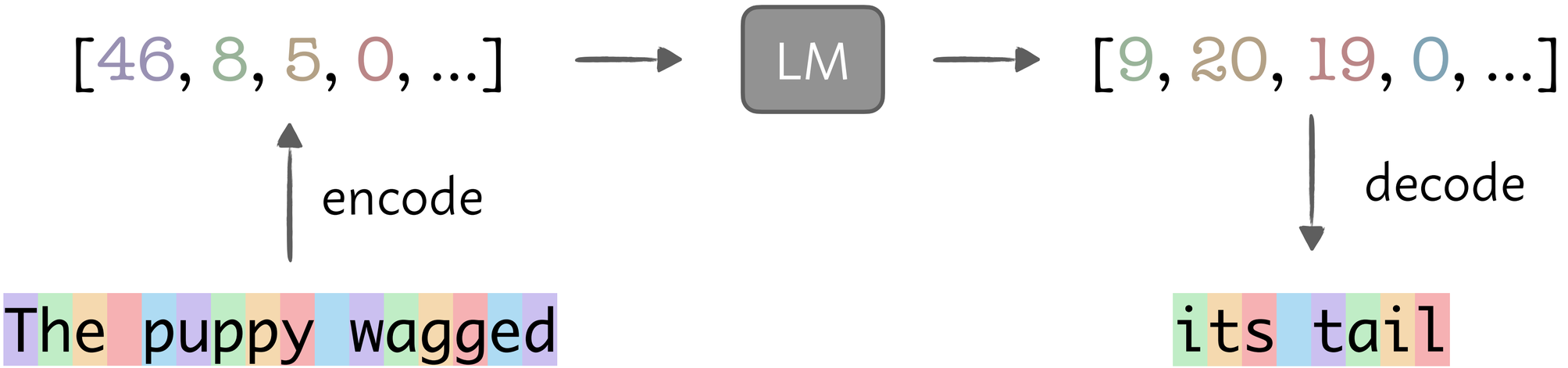
- 690 new words added in Sep 2023: "rizz," "goated," "bussin'," "mid"

Still need a way to deal with unknown words

Character-level tokenization

$$V = \{a, b, c, \ldots, z, A, B, C, \ldots, Z\}$$

(plus spaces + punctuation?)



Character-level tokenization

✅ Pros

Small vocabulary size

Complete coverage of input

Direct observation of spelling

❌ Cons

Super long sequences

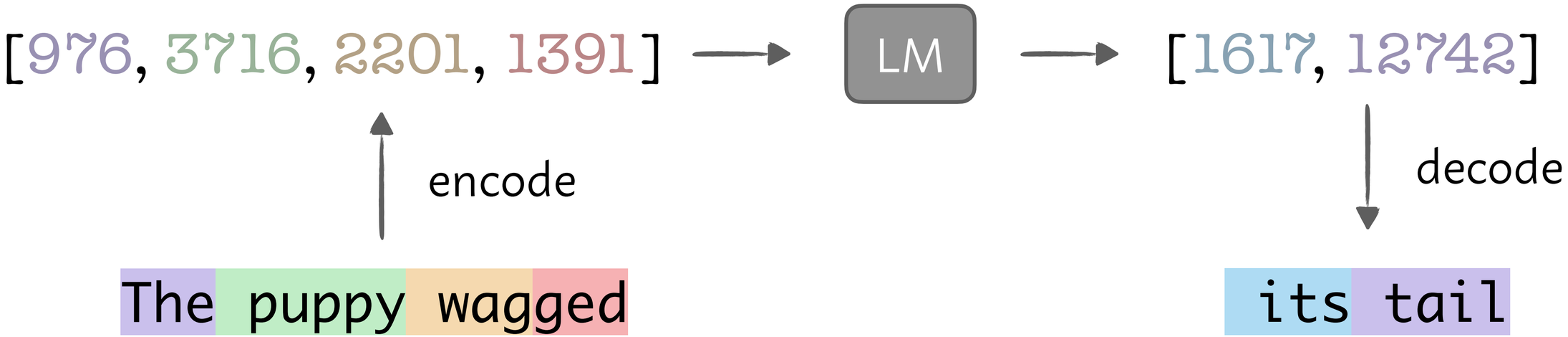
Difficult to learn over

Subword tokenization

How can we combine the high coverage of character-level representation with the efficiency of word-level representation?

Tokens are **subwords**, i.e., *parts* of words

Instead of defining the vocabulary a-priori, use *data* to tell us what our vocabulary should be



BPE: Byte Pair Encoding

(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)

BPE Algorithm

Required:

Training data D

Desired vocab size N

Algorithm:

1. Pretokenize D by splitting on whitespace
2. Initialize V as characters in D
3. Convert D into sequence of tokens (i.e., characters)
4. While $|V| < N$:
    a. Get counts of all bigrams (v_i, v_j) in D
    b. Merge most frequent pair into new token $v_n = v_i v_j$ where $n = |V| + 1$
    c. Replace all instances of $v_i v_j$ in D with v_n

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 characters

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 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

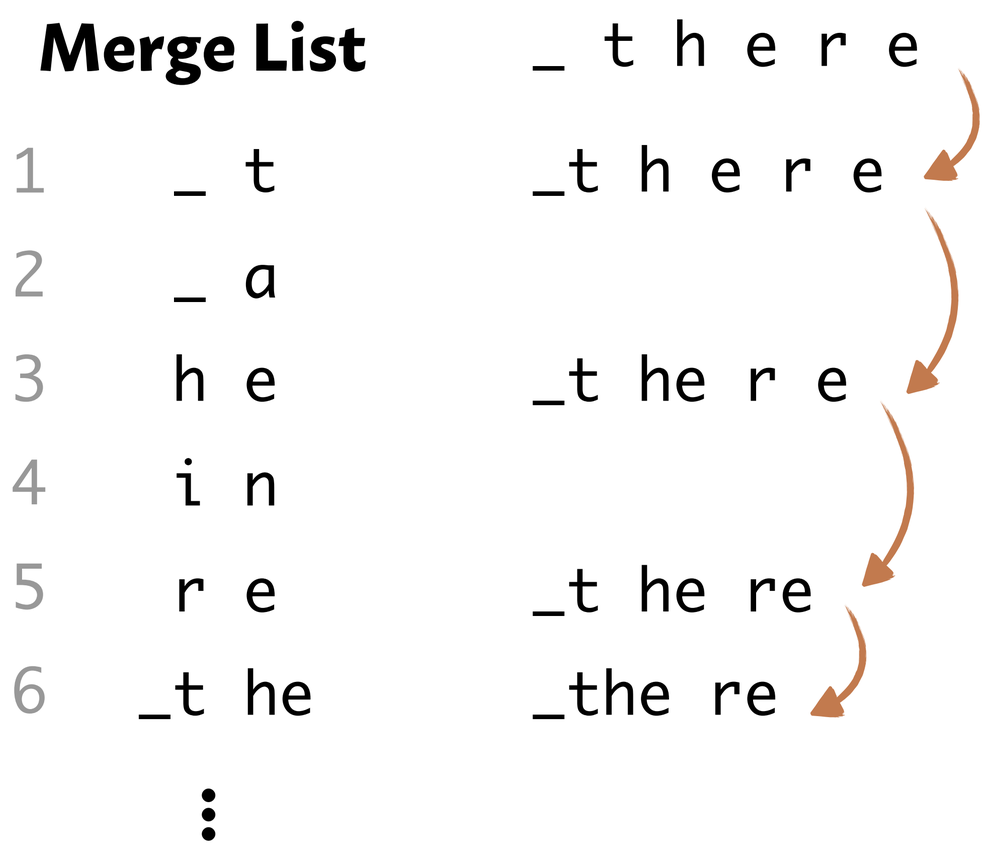
_the

⋮

until we reach desired vocab size $\mathscr{V}$

BPE Algorithm

To tokenize new text at test time, we split it into the characters and apply merge rules in order.



BPE: Examples

Given this BPE tokenizer, how would _the be tokenized?

Merge List

```
1   _ t
2  _t h
3   h e
```

Answer:

```
 _ t h e
_t h e
_th e
```

Merge List

```
1   _ t
2   h e
3  _t h
```

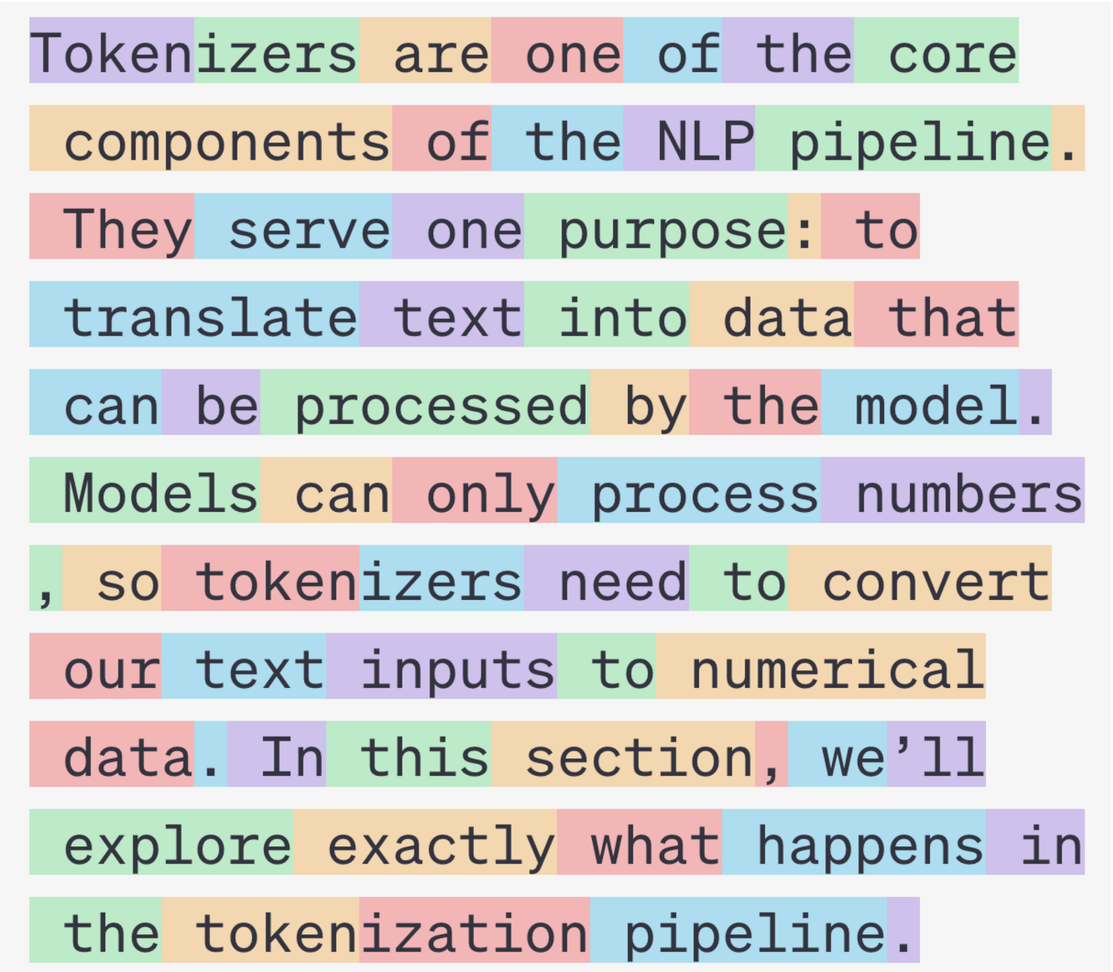
Answer:

```
 _ t h e
_t h e
_t he
```

ChatGPT's tokenizer

Tokenizers are one of the core components of the NLP pipeline. They serve one purpose: to translate text into data that can be processed by the model. Models can only process numbers, so tokenizers need to convert our text inputs to numerical data. In this section, we'll explore exactly what happens in the tokenization pipeline.

https://platform.openai.com/tokenizer



Subword tokenizers

✅ Pros

Everything can be represented with the vocabulary

Some shared representations

`wagged`, `wagging`

❌ Cons

No association between related tokens

`Run` ≠ `run` ≠ `RUN`

Learn the good, bad, & ugly in data

GPT-2 tokens[1]: `_RandomRedditor`, `_SolidGoldMagikarp`, `PsyNetMessage`

No direct observation of spelling

“Intermediate” tokens can be useless

`entucky` token is completely subsumed by `_Kentucky`

What could we do differently?

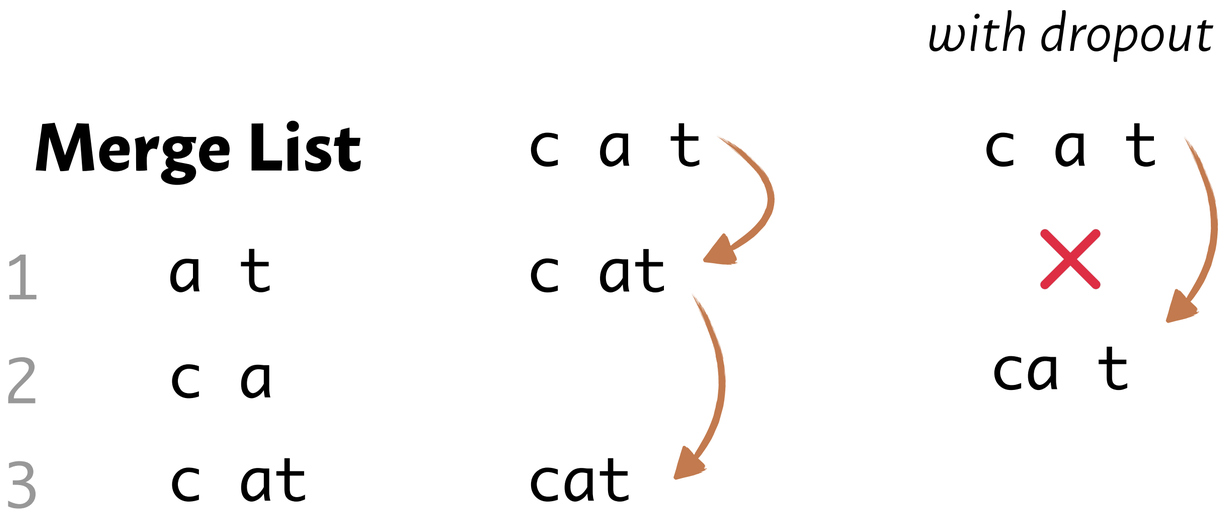
Non-deterministic tokenization

BPE always encodes the same text in the same way

There are actually many different ways to represent the same text with tokens in the vocabulary!

```
cat = {[cat], [c, at], [ca, t], [c,a,t]}
```

Achieve this by randomly drop some merges when encoding text



BPE-Dropout (Provilkov et al., 2020)

How to signal whitespace?

Instead of merging spaces into the beginning of words, use special "continue word" character

With whitespace: `[_Token, ization, _is, _cool]`

W/o whitespace: `[Token, ##ization, is, cool]`

Loses whitespace information

(especially problematic for code!)

```
from transformers import AutoTokenizer
tokenizer = AutoTokenizer.from_pretrained("openai-gpt")
```

✓ 0.4s

```
token_ids = tokenizer.encode("Tokenization is cool.")
print(token_ids)
print(tokenizer.decode(token_ids))
```

✓ 0.0s

```
[571, 2987, 26922, 544, 2548, 239]
tokenization is cool .
```

```
token_ids = tokenizer.encode("Tokenization      is     cool.")
print(token_ids)
print(tokenizer.decode(token_ids))
```

✓ 0.0s

```
[571, 2987, 26922, 544, 2548, 239]
tokenization is cool .
```

Byte-based vs. character-based?

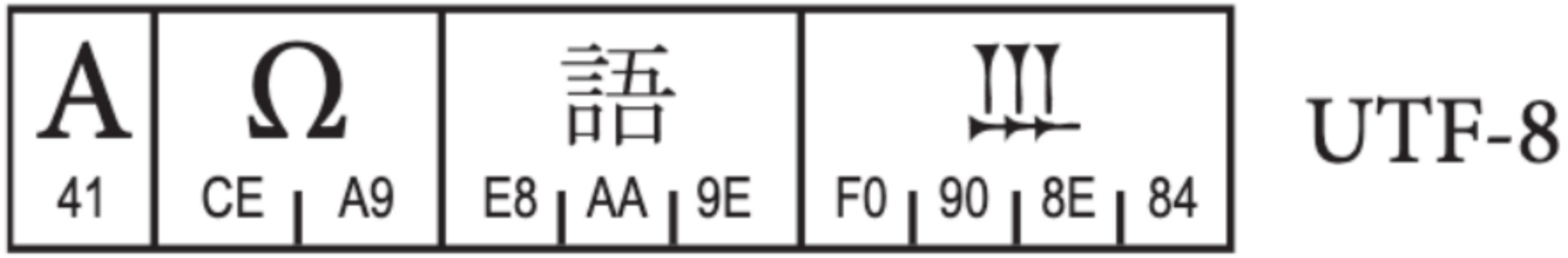
Originally, we presented BPE as having *characters* as the smallest unit

But there are *many* characters if you want to support...

- Character-based languages (e.g., שمঅ学ひ한กᏌЖ)
- Non-alphanumeric characters (e.g., 💀🫠😍)

Instead, use UTF-8 to map all characters in Unicode to byte strings (of 1-4 bytes)

Initialize base vocab as the set of 256 bytes, instead of the English characters



Pretokenization decisions?

Recall: pretokenization sets limits on what boundaries our tokens can cross

One of the main free variables in tokenizers today

How should we pretokenize on…

- Digits? Consider: 10, 1000000, 5493747
- Consecutive spaces? Consider:
- Punctuation? Consider: !=, (), get., .get
- Newlines? Consider: ;\n, \nThe

```
loop {
    // Stop as soon as we have a big enough vocabulary
    if word_to_id.len() >= self.vocab_size {
        break;
    }

    let mut top: Merge = queue.pop().unwrap();
```

The impact of tokenization on arithmetic in frontier LLMs (Singh & Strouse, 2024)

Subwords → superwords?

Why restrict tokens to *parts* of whitespace-delimited words?

- Multi-word expressions (e.g., “on the other hand”, “depend on”)
- Crosslingual variation in # of words needed for the same meaning (“Mathelehrer” in German)
- Some languages don’t use whitespace at all!

Including superword tokens can improve encoding efficiency

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

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

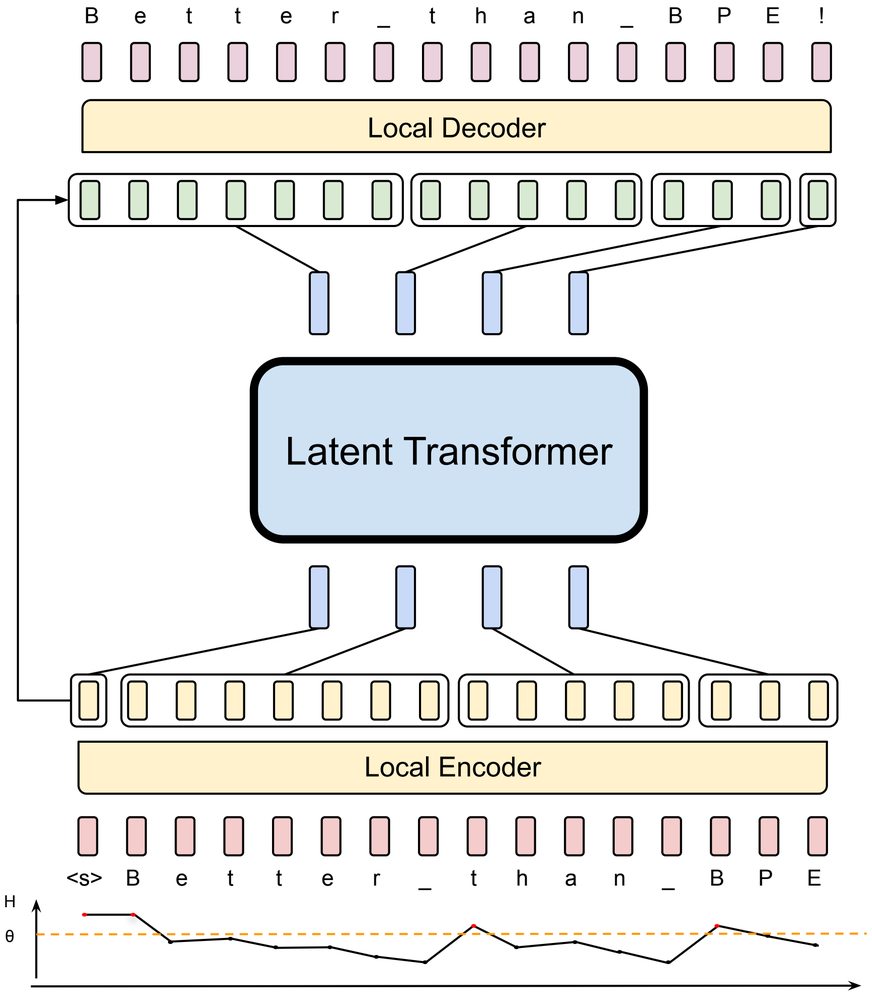
SuperBPE (Liu et al., 2025)

Tokenizer-free language models

See text in its “raw” form, as a sequence of bytes

Encode *all* characters with only 256 bytes

Learn token boundaries jointly with language modeling, or heuristically “patch” bytes into tokens



Dynamic Token Pooling (Nawrot et al., 2023)

MAGNET (Ahia et al., 2024)

Byte-Latent Transformers (Pagnoni et al., 2024)

Unigram tokenization

Intuition: initialize vocabulary as all substrings of all words, then repeatedly prune it until the desired vocab size is reached

Greedily optimize for high probability under a unigram LM

$$P(\texttt{[“breakfastish”]}) = \frac{1}{100} = 0.01$$

$$P(\texttt{[“breakfast”, “ish”]}) = P(\texttt{[“breakfast”]}) \cdot P(\texttt{[“ish”]}) = \frac{4}{100} \cdot \frac{10}{100} = 0.004$$

To encode new text, use the Viterbi algorithm to find the *optimal* encoding

Unigram tokenization (Kudo, 2018)