LLM LE3 VT2026

Data Processing

Fredrik Heintz
Dept. of Computer Science
Linköping University
fredrik.heintz@liu.se
@FredrikHeintz

Outline:

- Tokenization
- Data Processing Pipeline

LINKÖPING UNIVERSITY

Language modelling

- **Language modelling** is the task of predicting which word comes next in a sequence of words.
- More formally, given a sequence of words $w_1, \ldots, w_t$ we want to know the probability of the next word, w_{t+1}:

$$P(w_{t+1} \mid w_1, \ldots, w_t)$$

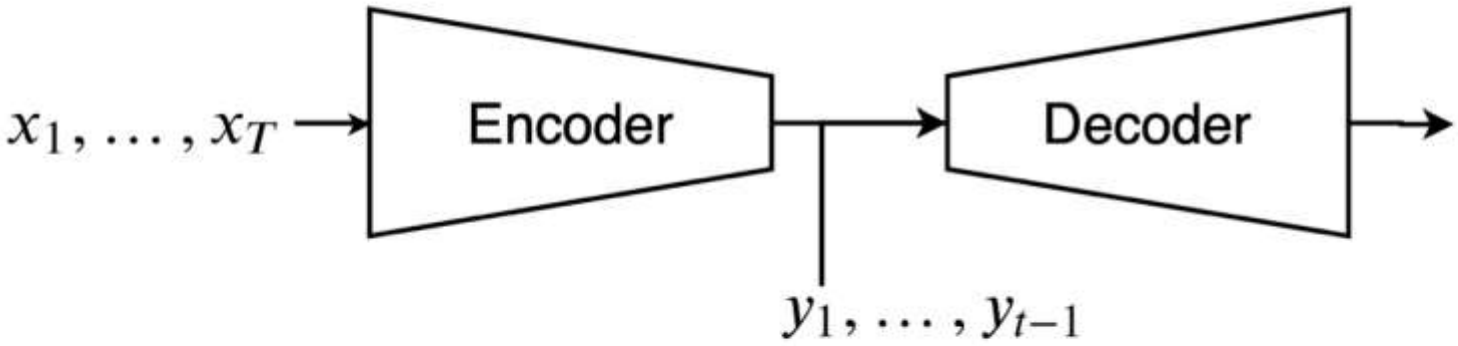
- We are assuming that w_{t+1} comes from a finite vocabulary V.

language models = classifiers

Funded by the European Union

Neurla Language Models

Input sequence: $x_1, \ldots, x_T$
Target sequence: $y_1, \ldots, y_T$



$P(Y_t = i)$

Or sometimes…

$P_\Theta(Y_t = i)$

Or sometimes…

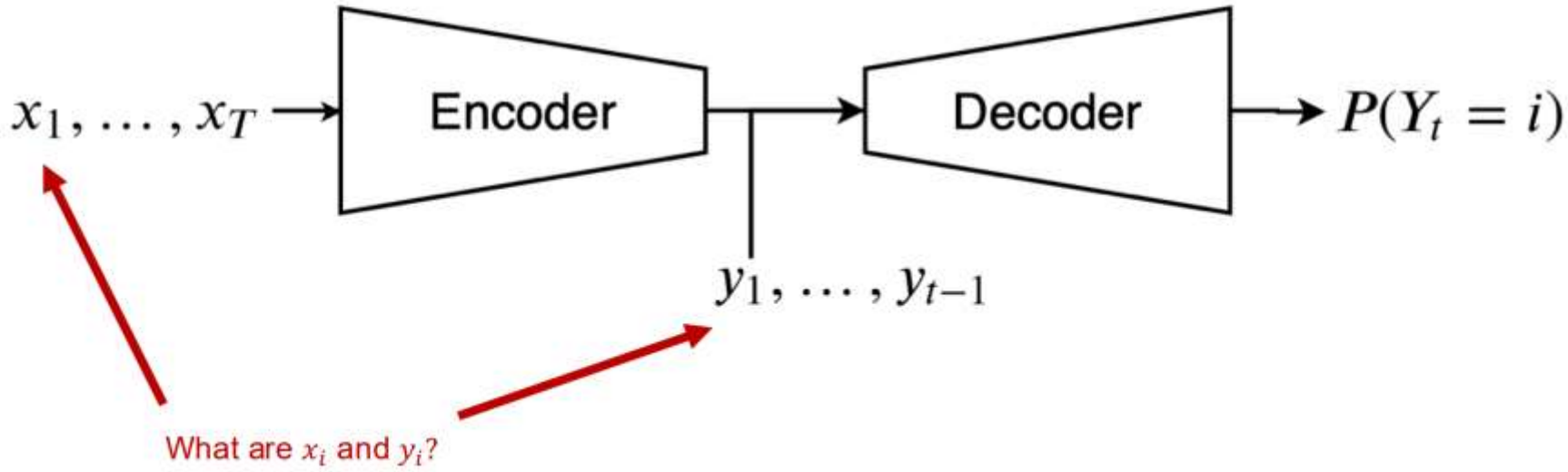
$P(Y_t = i | y_1, \ldots, y_{t-1}; x_1, \ldots, x_T; \Theta)$

Funded by the European Union

Neural Language Models

Input sequence: $x_1, \ldots, x_T$

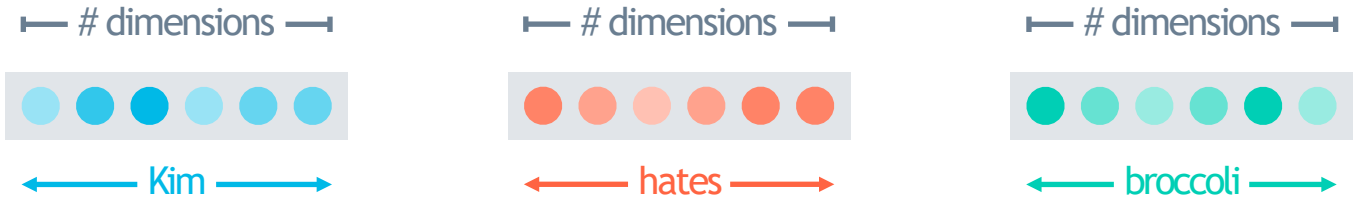
Target sequence: $y_1, \ldots, y_T$



Funded by the European Union

Word embeddings

To process words using neural networks, we need to represent them as vectors of numerical values.



Compared to one-hot vectors, **word embeddings**

- are shorter but dense
- support a useful notion of similarity
- can be learned from data

Funded by the European Union

Tokenisation

Funded by
the European Union

What is tokenisation?

- **Tokenisation** is the task of taking text (or code or music) and turning it into a sequence of discrete items, such as words or characters, called tokens.
- Tokenisation simplifies natural language processing by reducing unstructured text to more useful units.
- Tokenisation is the first step in mapping text to a numerical representation that computers can process.

Funded by the European Union

Words provide important signals

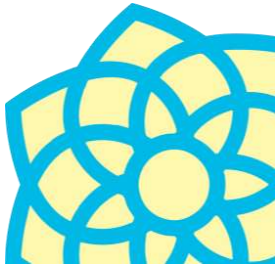
The gorgeously elaborate continuation of "The Lord of the Rings" trilogy is so huge that a column of words cannot adequately describe co-writer/director Peter Jackson's expanded vision of J.R.R. Tolkien's Middle-earth.

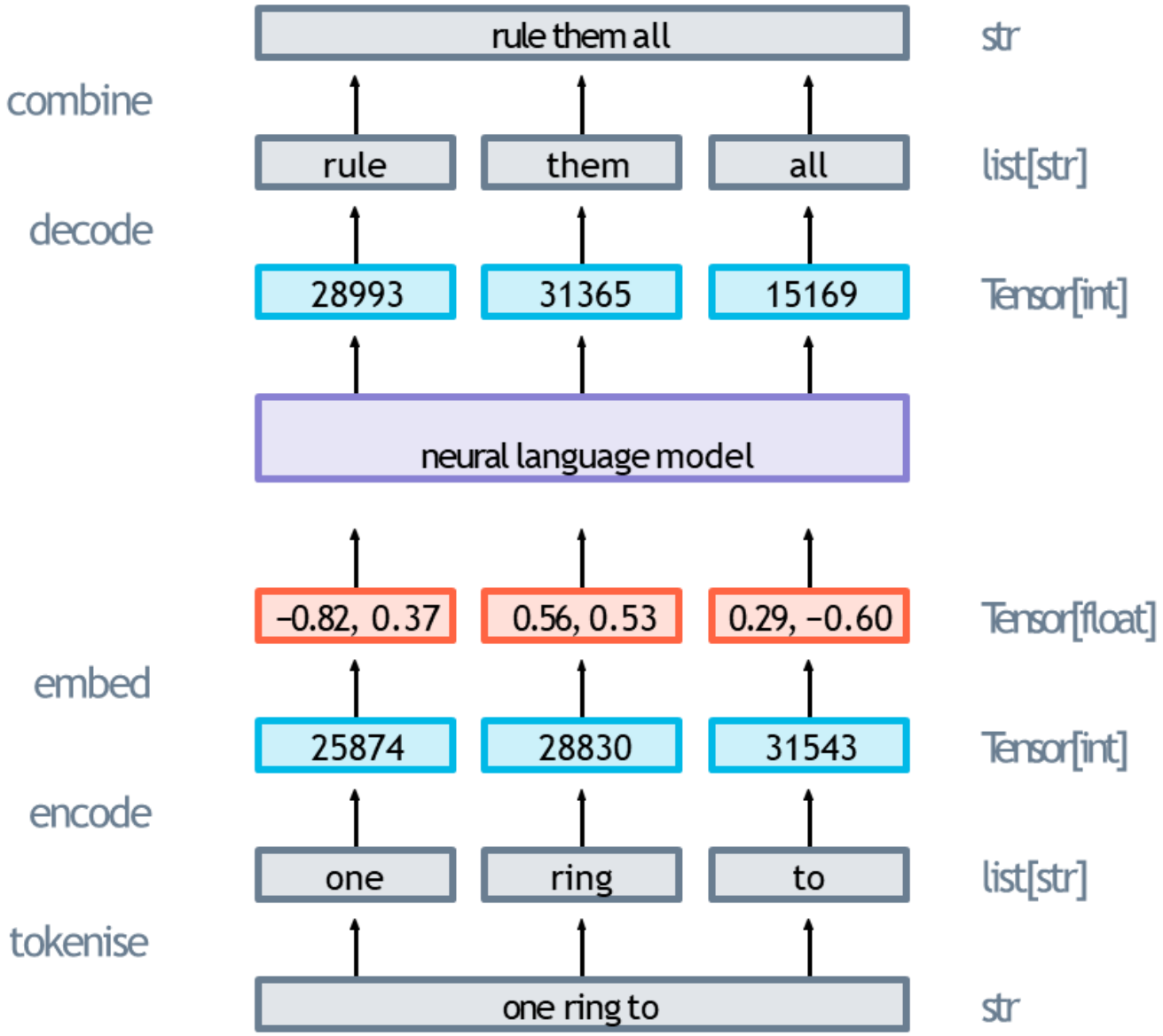
positive

… is a sour little movie at its core; an exploration of the emptiness that underlay the relentless gaiety of the 1920's, as if to stop would hasten the economic and global political turmoil that was to come.

negative

Funded by the European Union





Funded by the European Union

Whitespace tokenisation

```
# Tokenise text by splitting at whitespace def
tokenize(text: str) -> list[str]:
    return text.split()

# Create a vocabulary
vocab: set[str] = set(tokenize(text))
# {'cannot', 'huge', 'column', 'that', 'is', …}

# Create a string-to-ID mapping
stoid: dict[str, int] = {s: i for i, s in enumerate(vocab)}
# {'cannot': 0, 'huge': 1, 'column': 2, 'that': 3, 'is': 4, …}
```


Funded by the European Union

Whitespace tokenisation

The gorgeously elaborate continuation of "The Lord of the Rings" trilogy is so huge that a column of words cannot adequately describe co-writer/director Peter Jackson's expanded vision of J.R.R. Tolkien's Middle-earth.

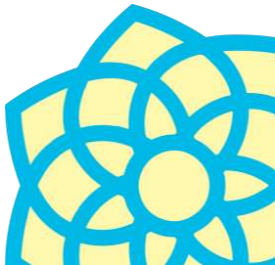
Regex-based tokenisation

The gorgeously elaborate continuation of " The Lord of the Rings " trilogy is so huge that a column of words cannot adequately describe co-writer / director Peter Jackson 's expanded vision of J. R. R. Tolkien 's Middle-earth .

```
re.findall(r"[A-Za-z]\.|\w+(?:-\w+)*|'\w+|[^\w\s]+", text)
```

single letters followed by a period	whole words, incl. hyphenated words	genitives ('s) and contractions ('ve)	punctuation, other non-word characters

Funded by the European Union

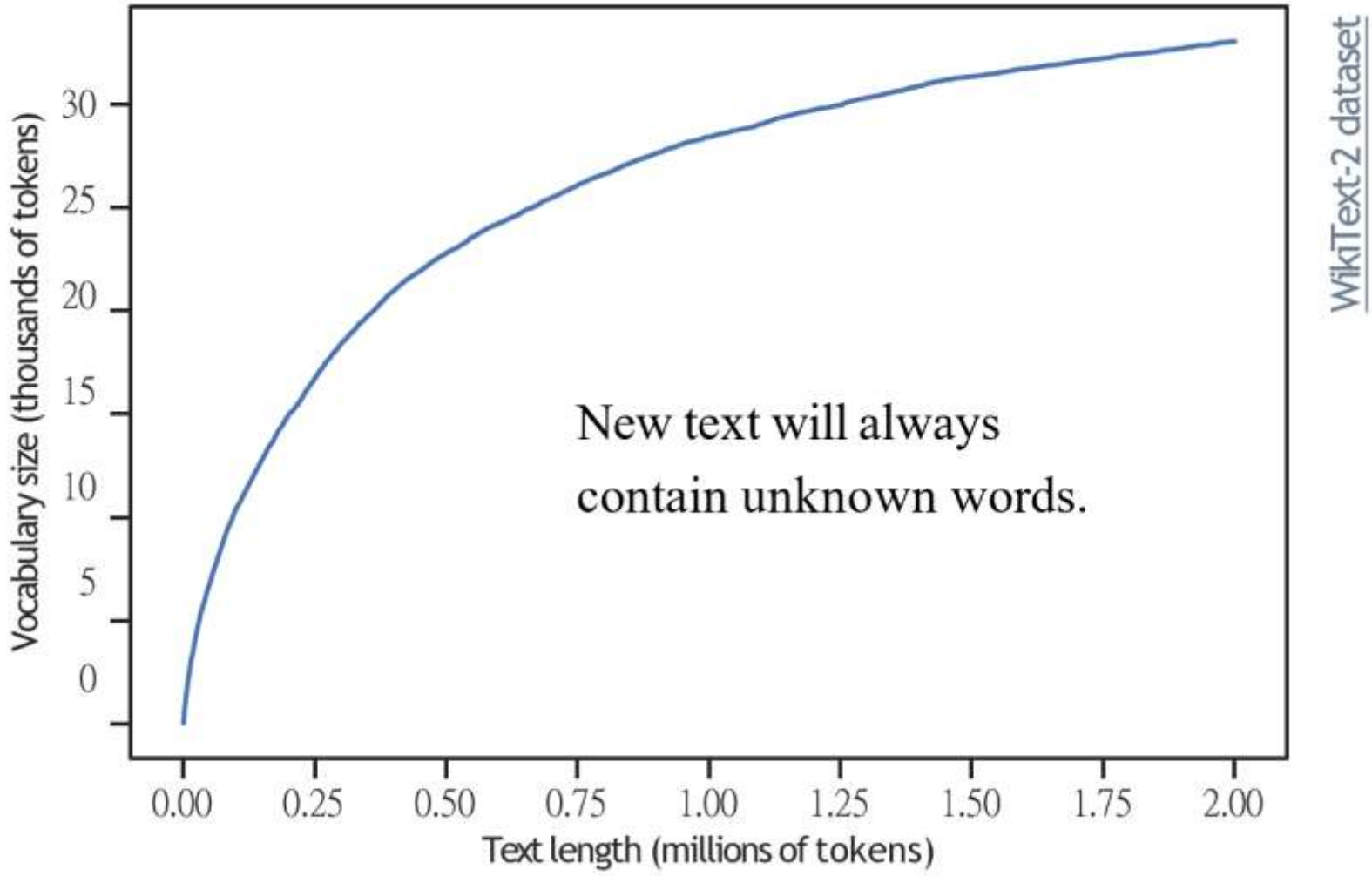


Text normalisation

- **Text normalisation** refers to the process of converting text into a more useful, standard form.
- Standard techniques include case normalisation, harmonisation of spelling variants, lemmatisation, and removing punctuation.
  Harmonisation: *color* →*colour*. Lemmatization: *runs*, *ran*, *running* →*run*
- Text normalisation was once a critical step in NLP tasks but is no longer as widely used today.

Funded by the European Union

The challenge of unknown words – Heaps' law



Funded by
the European Union

Dealing with unknown words

By Ena Music - Own work, CC BY-SA 4.0, Link

- **Step 1:** Build the vocabulary as usual.

  often combined with a frequency threshold

- **Step 2:** Augment the vocabulary with a special token, such as [UNK] to represent unknown words.

- **Step 3:** When processing new text, replace any out-of-vocabulary (oov) word with the special [UNK] token.

  The quokka is adorable. →*The [UNK] is adorable.* (Assuming *quokka* is oov.)

Funded by the European Union

But what is a word, anyway?

- There are many languages that do not adhere to the same concept of a "word" as English and Swedish.
- **Chinese** is written without spaces between characters. Identifying word boundaries is challenging.

  姚明进入总决赛 – "Yao Ming reaches the finals."
- **Inuktitut** allows entire sentences to be expressed as single words by combining multiple morphemes.

  tusaatsiarunnanngittualuujunga – "I cannot hear very well."

Funded by the European Union

Tokenization is language dependent

Table 1: Tokenizer comparisons between original LLaMA and Chinese LLaMA.

	Length	Content
Original Sentence	28	人工智能是计算机科学、心理学、哲学等学科融合的交叉学科。
Original Tokenizer	35	'_', '人', '工', '智', '能', '是', '计', '算', '机', '科', '学', '、', '心', '理', '学', '、', '0xE5', '0x93', '0xB2', '学', '等', '学', '科', '0xE8', '0x9E', '0x8D', '合', '的', '交', '0xE5', '0x8F', '0x89', '学', '科', '。'
Chinese Tokenizer	16	'_', '人工智能', '是', '计算机', '科学', '、', '心理学', '、', '哲学', '等', '学科', '融合', '的', '交叉', '学科', '。'

LLaMA tokenizer is **unfriendly** to Chinese

Yiming Cui. et.al. EFFICIENT AND EFFECTIVE TEXT ENCODING FOR CHINESE LLAMA AND ALPACA. https://arxiv.org/pdf/2304.08177.pdf

Funded by the European Union

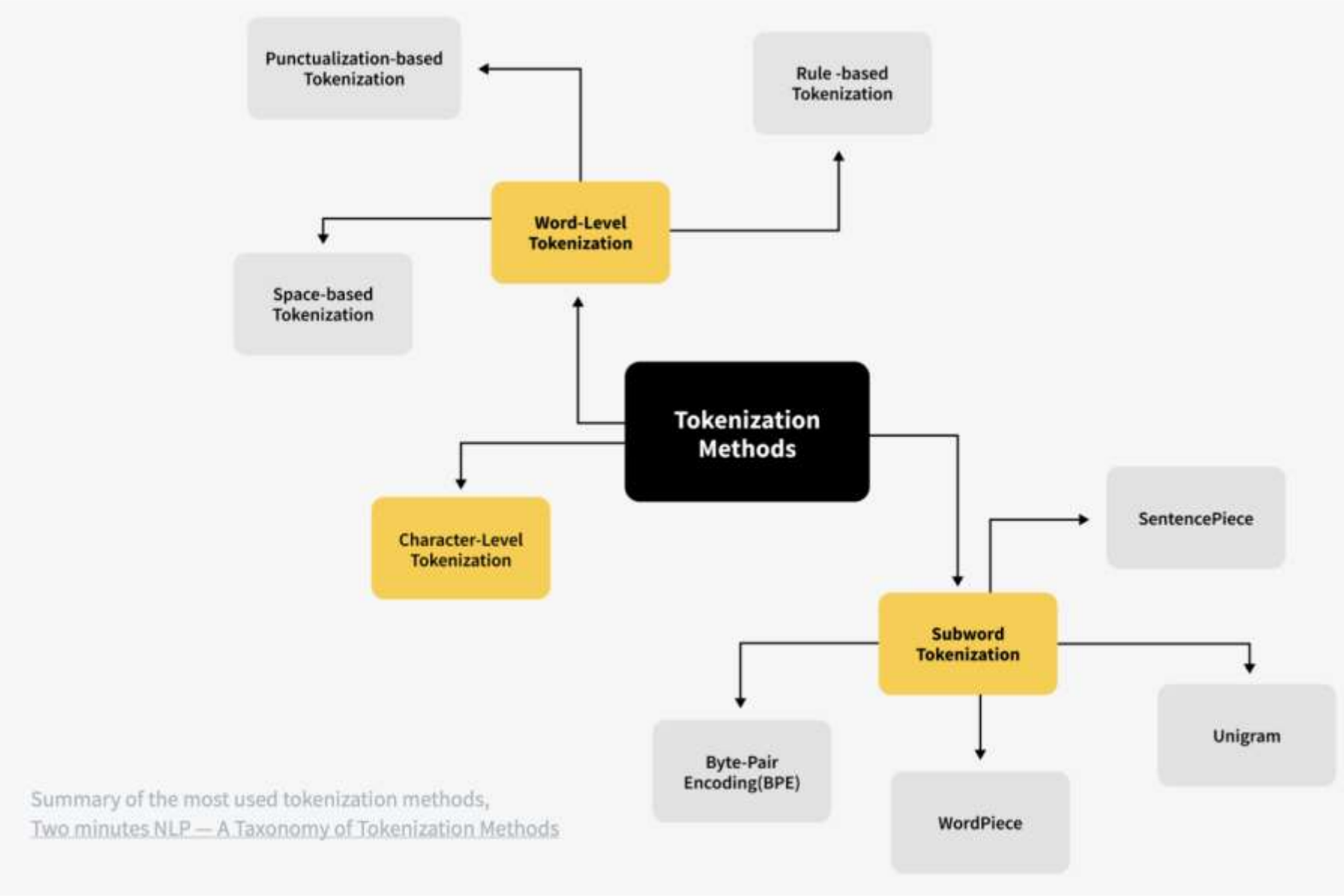
Target representations for tokenisation

- **Option 1:** Tokenise into words
  But: concept of "word" not universal; unknown words
- **Option 2:** Tokenise into individual characters
  But: may be too small a unit for learning
- **Option 3:** Tokenise into subwords
  Intuition: words are composed of morphemes

Let's tokenize: "A hippopotamus ate my homework."

Vocab Type	Example	Ex. length
character-level	['A', ' ', 'h', 'i', 'p', 'p', 'o', 'p', 'o', 't', 'a', 'm', 'u', 's', ' ', 'a', 't', 'e', ' ', 'm', 'y', ' ', 'h', 'o', 'm', 'e', 'w', 'o', 'r', 'k', '.']	31
subword-level	['A', 'hip', '##pop', '##ota', '##mus', 'ate', 'my', 'homework', '.']	9
word-level	['A', 'hippopotamus', 'ate', 'my', 'homework']	5

Funded by the European Union



Summary of the most used tokenization methods,
Two minutes NLP — A Taxonomy of Tokenization Methods

Funded by the European Union

Tokenization Methods	Word-based tokenization	Character-based tokenization	Subword-based tokenization
Example Tokenizers	Space tokenization (split sentences by space); rule-based tokenization (e.g. Moses, spaCy)	Character tokenization (simply tokenize on every character)	Byte-Pair Encoding (BPE); WordPiece; SentencePiece; Unigram (tokenizing by parts of a word vs. the entirety of a word; see table above)
Considerations	• **Downside:** Generates a very large vocabulary leading to a huge embedding matrix as the input and output layer; large number of out-of-vocabulary (OOV) tokens; and different meanings of very similar words • Transformer models normally have a vocabulary of less than 50,000 words, especially if they are trained only on a single language	• Lead to much smaller vocabulary; no OOV (out of vocabulary) tokens since every word can be assembled from individual characters • **Downside:** Generates very long sequences and less meaningful individual tokens, making it harder for the model to learn meaningful input representations. However, if character-based tokenization is used on non-English language, a single character could be quite information rich (like "mountain" in Mandarin).	• Subword-based tokenization methods follow the principle that frequently used words should not be split into smaller subwords, but rare words should be decomposed into meaningful subwords • **Benefit:** Solves the downsides faced by word-based tokenization and character-based tokenization and achieves both reasonable vocabulary size with meaningful learned context-independent representations.

LooongLLaVA
LGUer

Funded by the European Union

Subword-based Tokenization Methods

- **Byte-Pair Encoding** [Gage 1994]
  - Originally used in machine translation
- **WordPiece**
- **Unigram**
- **SentencePiece**

Subword-based Tokenization Methods	Byte-Pair Encoding (BPE)	WordPiece	Unigram	SentencePiece
Description	One of the most popular subword tokenization algorithms. The Byte-Pair-Encoding works by starting with characters, while merging those that are the most frequently seen together, thus creating new tokens. It then works iteratively to build new tokens out of the most frequent pairs it sees in a corpus. BPE is able to build words it has never seen by using multiple subword tokens, and thus requires smaller vocabularies, with less chances of having "unk" (unknown) tokens.	Very similar to BPE. The difference is that WordPiece does not choose the highest frequency symbol pair, but the one that maximizes the likelihood of the training data once added to the vocabulary (evaluates what it loses by merging two symbols to ensure it's worth it)	In contrast to BPE / WordPiece, Unigram initializes its base vocabulary to a large number of symbols and progressively trims down each symbol to obtain a smaller vocabulary. It is often used together with SentencePiece.	The left 3 tokenizers assume input text uses spaces to separate words, and therefore are not usually applicable to languages that don't use spaces to separate words (e.g. Chinese). SentencePiece treats the input as a raw input stream, thus including the space in the set of characters to use. It then uses the BPE / Unigram algorithm to construct the appropriate vocabulary.
Considerations	BPE is particularly useful for handling rare and out-of-vocabulary words since it can generate subwords for new words based on the most common character sequences. Downside: BPE can result in subwords that do not correspond to linguistically meaningful units.	WordPiece can be particularly useful for languages where the meaning of a word can depend on the context in which it appears.	Unigram tokenization is particularly useful for languages with complex morphology and can generate subwords that correspond to linguistically meaningful units. However, unigram tokenization can struggle with rare and out-of-vocabulary words.	SentencePiece can be particularly useful for languages where the meaning of a word can depend on the context in which it appears.

Funded by the European Union

Byte-Pair Encoding (BPE) [Gage 1994]

- **Byte Pair Encoding (BPE)** is an algorithm for learning subword tokens from text.

- **Step 1:** Encode the text into a sequence of bytes. Initialise the token vocabulary with all single bytes.
- **Step 2:** Create a new token by merging the most frequent pair of consecutive tokens. Add the new token to the vocabulary.
- Repeat the previous step as long as the token vocabulary does not exceed a predefined maximum size.

Sennrich et al. (2016)

Funded by the European Union

The Unicode Standard

- **Unicode** is a text encoding standard designed to support text from all the world's writing systems (that can be digitised).
- Version 16.0 supports 154,998 characters from 168 scripts.
- For backwards compatibility, the first 128 codepoints of Unicode are the same as ASCII.

Funded by the European Union

	000	001	002	003	004	005	006	007
0	NUL	DLE	SP	0	@	P	`	p
1	SOH	DC1	!	1	A	Q	a	q
2	STX	DC2	"	2	B	R	b	r
3	ETX	DC3	#	3	C	S	c	s
4	EOT	DC4	$	4	D	T	d	t
5	ENQ	NAK	%	5	E	U	e	u
6	ACK	SYN	&	6	F	V	f	v
7	BEL	ETB	'	7	G	W	g	w
8	BS	CAN	(	8	H	X	h	x
9	HT	EM	)	9	I	Y	i	y
A	LF	SUB	*	:	J	Z	j	z
B	VT	ESC	+	;	K	[	k	{
C	FF	FS	,	<	L	\	l	|
D	CR	GS	-	=	M	]	m	}
E	SO	RS	.	>	N	^	n	~
F	SI	US	/	?	O	_	o	DEL

Various signs

0900 ँ DEVANAGARI SIGN INVERTED CANDRABINDU
= vaidika adhomukha candrabindu
0901 ं DEVANAGARI SIGN CANDRABINDU
= anunasika
→ 0310 ◌̐ combining candrabindu
0902 ं DEVANAGARI SIGN ANUSVARA
= bindu
0903 ः DEVANAGARI SIGN VISARGA

Independent vowels

0904 ऄ DEVANAGARI LETTER SHORT A
• used for short e in Awadhi
• also used in Devanagari transliterations of some South Indian and Kashmiri languages by a publisher in Lucknow
0905 अ DEVANAGARI LETTER A
0906 आ DEVANAGARI LETTER AA
0907 इ DEVANAGARI LETTER I
0908 ई DEVANAGARI LETTER II
0909 उ DEVANAGARI LETTER U
090A ऊ DEVANAGARI LETTER UU
090B ऋ DEVANAGARI LETTER VOCALIC R
090C ऌ DEVANAGARI LETTER VOCALIC L
090D ऍ DEVANAGARI LETTER CANDRA E
090E ऎ DEVANAGARI LETTER SHORT E
• Kashmiri, Bihari languages
• also used for transcribing Dravidian short e
090F ए DEVANAGARI LETTER E
0910 ऐ DEVANAGARI LETTER AI
0911 ऑ DEVANAGARI LETTER CANDRA O
0912 ऒ DEVANAGARI LETTER SHORT O
• Kashmiri, Bihari languages
• also used for transcribing Dravidian short o
0913 ओ DEVANAGARI LETTER O
0914 औ DEVANAGARI LETTER AU

Consonants

0915 क DEVANAGARI LETTER KA
0916 ख DEVANAGARI LETTER KHA
0917 ग DEVANAGARI LETTER GA
0918 घ DEVANAGARI LETTER GHA
0919 ङ DEVANAGARI LETTER NGA
091A च DEVANAGARI LETTER CA
091B छ DEVANAGARI LETTER CHA
091C ज DEVANAGARI LETTER JA
091D झ DEVANAGARI LETTER JHA
091E ञ DEVANAGARI LETTER NYA
091F ट DEVANAGARI LETTER TTA
0920 ठ DEVANAGARI LETTER TTHA
0921 ड DEVANAGARI LETTER DDA
0922 ढ DEVANAGARI LETTER DDHA
0923 ण DEVANAGARI LETTER NNA
0924 त DEVANAGARI LETTER TA
0925 थ DEVANAGARI LETTER THA
0926 द DEVANAGARI LETTER DA
0927 ध DEVANAGARI LETTER DHA
0928 न DEVANAGARI LETTER NA
0929 ऩ DEVANAGARI LETTER NNNA
• for transcribing Dravidian alveolar n
= 0928 न 093C ◌़
092A प DEVANAGARI LETTER PA
092B फ DEVANAGARI LETTER PHA
092C ब DEVANAGARI LETTER BA
092D भ DEVANAGARI LETTER BHA
092E म DEVANAGARI LETTER MA
092F य DEVANAGARI LETTER YA
0930 र DEVANAGARI LETTER RA
0931 ऱ DEVANAGARI LETTER RRA
• for transcribing Dravidian alveolar r
• half form is represented as "Eyelash RA"
= 0930 र 093C ◌़
0932 ल DEVANAGARI LETTER LA
0933 ळ DEVANAGARI LETTER LLA
0934 ऴ DEVANAGARI LETTER LLLA
• for transcribing Dravidian l
= 0933 ळ 093C ◌़
0935 व DEVANAGARI LETTER VA
0936 श DEVANAGARI LETTER SHA
0937 ष DEVANAGARI LETTER SSA
0938 स DEVANAGARI LETTER SA
0939 ह DEVANAGARI LETTER HA

Dependent vowel signs

These dependent vowel signs are used in Kashmiri and in the Bihari languages (Bhojpuri, Magadhi, and Maithili).

093A ऺ DEVANAGARI VOWEL SIGN OE
093B ऻ DEVANAGARI VOWEL SIGN OOE

Various signs

093C ◌़ DEVANAGARI SIGN NUKTA
• for extending the alphabet to new letters
093D ऽ DEVANAGARI SIGN AVAGRAHA

Dependent vowel signs

093E ा DEVANAGARI VOWEL SIGN AA
093F ि DEVANAGARI VOWEL SIGN I
• stands to the left of the consonant
0940 ी DEVANAGARI VOWEL SIGN II
0941 ु DEVANAGARI VOWEL SIGN U
0942 ू DEVANAGARI VOWEL SIGN UU
0943 ृ DEVANAGARI VOWEL SIGN VOCALIC R
0944 ॄ DEVANAGARI VOWEL SIGN VOCALIC RR
0945 ॅ DEVANAGARI VOWEL SIGN CANDRA E
= candra
0946 ॆ DEVANAGARI VOWEL SIGN SHORT E
• Kashmiri, Bihari languages
• also used for transcribing Dravidian short e
0947 े DEVANAGARI VOWEL SIGN E
0948 ै DEVANAGARI VOWEL SIGN AI
0949 ॉ DEVANAGARI VOWEL SIGN CANDRA O
094A ॊ DEVANAGARI VOWEL SIGN SHORT O
• Kashmiri, Bihari languages
• also used for transcribing Dravidian short o
094B ो DEVANAGARI VOWEL SIGN O
094C ौ DEVANAGARI VOWEL SIGN AU

Virama

094D ् DEVANAGARI SIGN VIRAMA
= halant (the preferred Hindi name)
• suppresses inherent vowel

Dependent vowel signs

094E ॎ DEVANAGARI VOWEL SIGN PRISHTHAMATRA E
• character has historic use only
• combines with E to form AI, with AA to form O, and with O to form AU

Funded by the European Union

Encoding text into bytes

- Encoding all (more than 1 million) Unicode characters into bytes requires more than one byte per character.
- **UTF-8 (8-bit Unicode Transformation Format)** is the most widely used encoding scheme for Unicode.
- It uses a variable-width encoding of 1-4 bytes per character.

  The first byte indicates how many additional bytes are part of the character.

Funded by the European Union

Encoding text into bytes

Einu sinni deildu norðanvindurinn og sólin um, kvort þeirra væri sterkara.

74 Unicode characters

E	i	n	u		s	i	n	n	i		d	e	i	l	d	u		n	o	r	ð		a
69	105	110	117	32	115	105	110	110	105	32	100	101	105	108	100	117	32	110	111	114	195	176	97

n	v	i	n	d	u	r	i	n	n		o	g		s	ó	l	i	n	n		u	m
110	118	105	110	100	117	114	105	110	110	32	111	103	32	115	195	179	108	105	110	110	32	117	109

,		k	v	o	r	t		þ	e	i	r	r	a		v	æ	r	i		s	t
44	32	107	118	111	114	116	32	195	190	101	105	114	114	97	32	118	195	166	114	105	32	115	116

e	r	k	a	r	a	.
101	114	107	97	114	97	46

78 bytes in UTF-8

Funded by the European Union

Byte-Pair Encoding – Example

```
The North Wind and the Sun were disputing which was the stronger,
when a traveler came along wrapped in a warm cloak. They agreed that
the one who first succeeded in making the traveler take his cloak
off should be considered stronger than the other.
```

Vocabulary (without single bytes)

Token ID	Token
256	
257	
258	
259	
260	

Pair counts

Token pair	Count
e + SPACE	11
SPACE + t	10
h + e	9
t + h	9
d + SPACE	7

Funded by the European Union

Byte-Pair Encoding – Example

The North Wind and the Sun were disputing which was the stronger, when a traveler came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveler take his cloak off should be considered stronger than the other.

Vocabulary (without single bytes)

Token ID	Token
256	[e]
257	
258	
259	
260	

Pair counts

Token pair	Count
e + SPACE	11
SPACE + t	10
h + e	9
t + h	9
d + SPACE	7

Funded by the European Union

Byte-Pair Encoding – Example

```
Th[e ]North Wind and th[e ]Sun wer[e ]disputing which was th[e ]stronger,
when a traveler cam[e ]along wrapped in a warm cloak. They agreed that
th[e ]on[e ]who first succeeded in making th[e ]traveler tak[e ]his cloak
off should b[e ]considered stronger than th[e ]other.
```

Vocabulary (without single bytes)

Token ID	Token
256	[e]
257	
258	
259	
260	

Pair counts

Token pair	Count
t + h	9
SPACE + t	9
d + SPACE	7
e + r	7
h + [e]	6

Funded by the European Union

Byte-Pair Encoding – Example

```
Th[e ]North Wind and th[e ]Sun wer[e ]disputing which was
th[e ]stronger, when a traveler cam[e ]along wrapped in a warm
cloak. They agreed that th[e ]on[e ]who first succeeded in making
th[e ]traveler tak[e ]his cloak off should b[e ]considered stronger
than th[e ]other.
```

Vocabulary (without single bytes)

Token ID	Token
256	[e]
257	[th]
258	
259	
260	

Pair counts

Token pair	Count
t + h	9
SPACE + t	9
d + SPACE	7
e + r	7
h + [e]	6

Funded by the European Union

Byte-Pair Encoding – Example

```
Th[e ]Nor[th] Wind and [th][e ]Sun wer[e ]disputing which was
[th][e ]stronger, when a traveler cam[e ]along wrapped in a warm
cloak. [Th]ey agreed [th]at [th][e ]on[e ]who first succeeded in making
[th][e ]traveler tak[e ]his cloak off should b[e ]considered stronger
[th]an [th][e ]other.
```

Vocabulary (without single bytes)

Token ID	Token
256	[e]
257	[th]
258	
259	
260	

Pair counts

Token pair	Count
d + SPACE	7
SPACE + [th]	7
e + r	7
SPACE + w	6
i + n	5

Funded by the European Union

Byte-Pair Encoding – Example

```
Th[e ]Nor[th] Wind and [th][e ]Sun wer[e ]disputing which was
[th][e ]stronger, when a traveler cam[e ]along wrapped in a warm
cloak. [Th]ey agreed [th]at [th][e ]on[e ]who first succeeded in making
[th][e ]traveler tak[e ]his cloak off should b[e ]considered stronger
[th]an [th][e ]other.
```

Vocabulary (without single bytes)

Token ID	Token
256	[e]
257	[th]
258	[d]
259	
260	

Pair counts

Token pair	Count
d + SPACE	7
SPACE + [th]	7
e + r	7
SPACE + w	6
i + n	5

Funded by the European Union

Byte-Pair Encoding – Example

```
Th[e ]Nor[th] Wind and [th][e ]Sun wer[e ]disputing which was
[th][e ]stronger, when a traveler cam[e ]along wrapped in a warm
cloak. [Th]ey agreed [th]at [th][e ]on[e ]who first succeeded in making
[th][e ]traveler tak[e ]his cloak off should b[e ]considered stronger
[th]an [th][e ]other.
```

Vocabulary (without single bytes)

Token ID	Token
256	[e]
257	[th]
258	[d]
259	
260	

Pair counts

Token pair	Count
e + r	7
SPACE + w	6
i + n	5
[th] + [e]	5
n + SPACE	5

Funded by the European Union

Byte-Pair Encoding – Example

```
Th[e ]Nor[th] Wind and [th][e ]Sun wer[e ]disputing which was
[th][e ]stronger, when a traveler cam[e ]along wrapped in a warm
cloak. [Th]ey agreed [th]at [th][e ]on[e ]who first succeeded in making
[th][e ]traveler tak[e ]his cloak off should b[e ]considered stronger
[th]an [th][e ]other.
```

Vocabulary (without single bytes)

Token ID	Token
256	[e]
257	[th]
258	[d]
259	
260	

Pair counts

Token pair	Count
e + r	7
SPACE + w	6
i + n	5
[th] + [e]	5
n + SPACE	5

Funded by the European Union

Byte-Pair Encoding – Example

```
Th[e ]Nor[th] Wind and [th][e ]Sun wer[e ]disputing which was
[th][e ]stronger, when a traveler cam[e ]along wrapped in a warm
cloak. [Th]ey agreed [th]at [th][e ]on[e ]who first succeeded in making
[th][e ]traveler tak[e ]his cloak off should b[e ]considered stronger
[th]an [th][e ]other.
```

Vocabulary (without single bytes)

Token ID	Token
256	[e]
257	[th]
258	[d]
259	[er]
260	

Pair counts

Token pair	Count
SPACE + w	6
i + n	5
[th] + [e]	5
n + SPACE	5
n + g	5

Funded by the European Union

Byte-Pair Encoding – Example

```
Th[e ]Nor[th] Wind and [th][e ]Sun wer[e ]disputing which was
[th][e ]stronger, when a traveler cam[e ]along wrapped in a warm
cloak. [Th]ey agreed [th]at [th][e ]on[e ]who first succeeded in making
[th][e ]traveler tak[e ]his cloak off should b[e ]considered stronger
[th]an [th][e ]other.
```

Vocabulary (without single bytes)

Token ID	Token
256	[e]
257	[th]
258	[d]
259	[er]
260	

Pair counts

Token pair	Count
SPACE + w	6
i + n	5
[th] + [e]	5
n + SPACE	5
n + g	5

Funded by the European Union

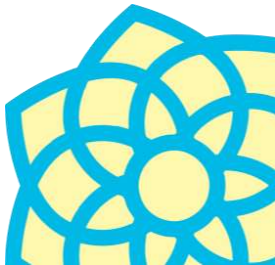
Byte-Pair Encoding – Example

```
Th[e ]Nor[th] Wind and [th][e ]Sun wer[e ]disputing which was
[th][e ]stronger, when a traveler cam[e ]along wrapped in a warm
cloak. [Th]ey agreed [th]at [th][e ]on[e ]who first succeeded in making
[th][e ]traveler tak[e ]his cloak off should b[e ]considered stronger
[th]an [th][e ]other.
```

Vocabulary (without single bytes)

Token ID	Token
256	[e]
257	[th]
258	[d]
259	[er]
260	[w]

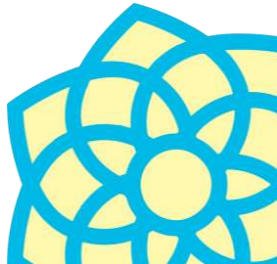
Funded by the European Union



Some comments on BPE

- The tokens obtained using BPE match varying spans of source text, from single characters to whole words and beyond.
- The tokens are not guaranteed to have any apparent linguistic meaning, but often resemble words or morphemes.
  BPE = “poor man’s morphology”
- BPE solves the problem with unknown words: Every text can be tokenised; in the worst case, it is tokenised as bytes.

Funded by the European Union



Tiktokenizer

o200k_base

The North Wind and the Sun were disputing which was the stronger, when a traveler came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveler take his cloak off should be considered stronger than the other.

Token count
49

The North Wind and the Sun were disputing which was t he stronger, when a traveler came along wrapped in a warm cloak. They agreed that the one who first succee ded in making the traveler take his cloak off should be considered stronger than the other.

976, 7180, 28551, 326, 290, 11628, 1504, 28301, 289, 1118, 673, 290, 26929, 11, 1261, 261, 72819, 5831, 42 51, 31831, 306, 261, 9144, 152842, 13, 3164, 12863, 4 84, 290, 1001, 1218, 1577, 53434, 306, 4137, 290, 728 19, 2304, 1232, 152842, 1277, 1757, 413, 9474, 26929, 1572, 290, 1273, 13

Show whitespace

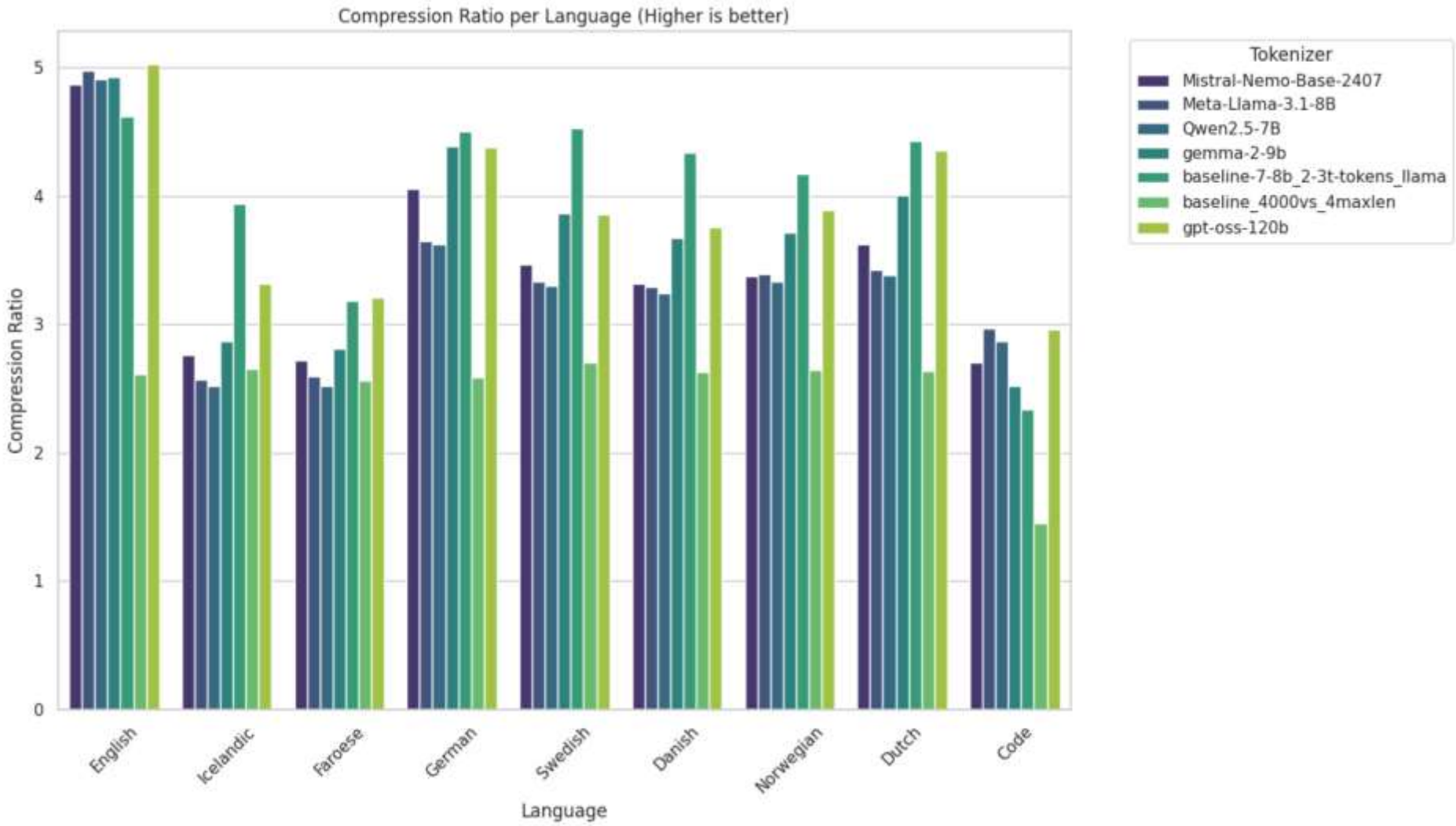
Built by dqbd. Created with the generous help from Diagram.

Tokenisation in language models

Model	Release year	Tokenisation method	Vocabulary size
BERT	2018	WordPiece	30 K
GPT-2	2019	BPE	50 K
GPT-3.5	2022	BPE	100 K
GPT-4o	2024	BPE	200 K
Llama 3	2024	BPE	128 K

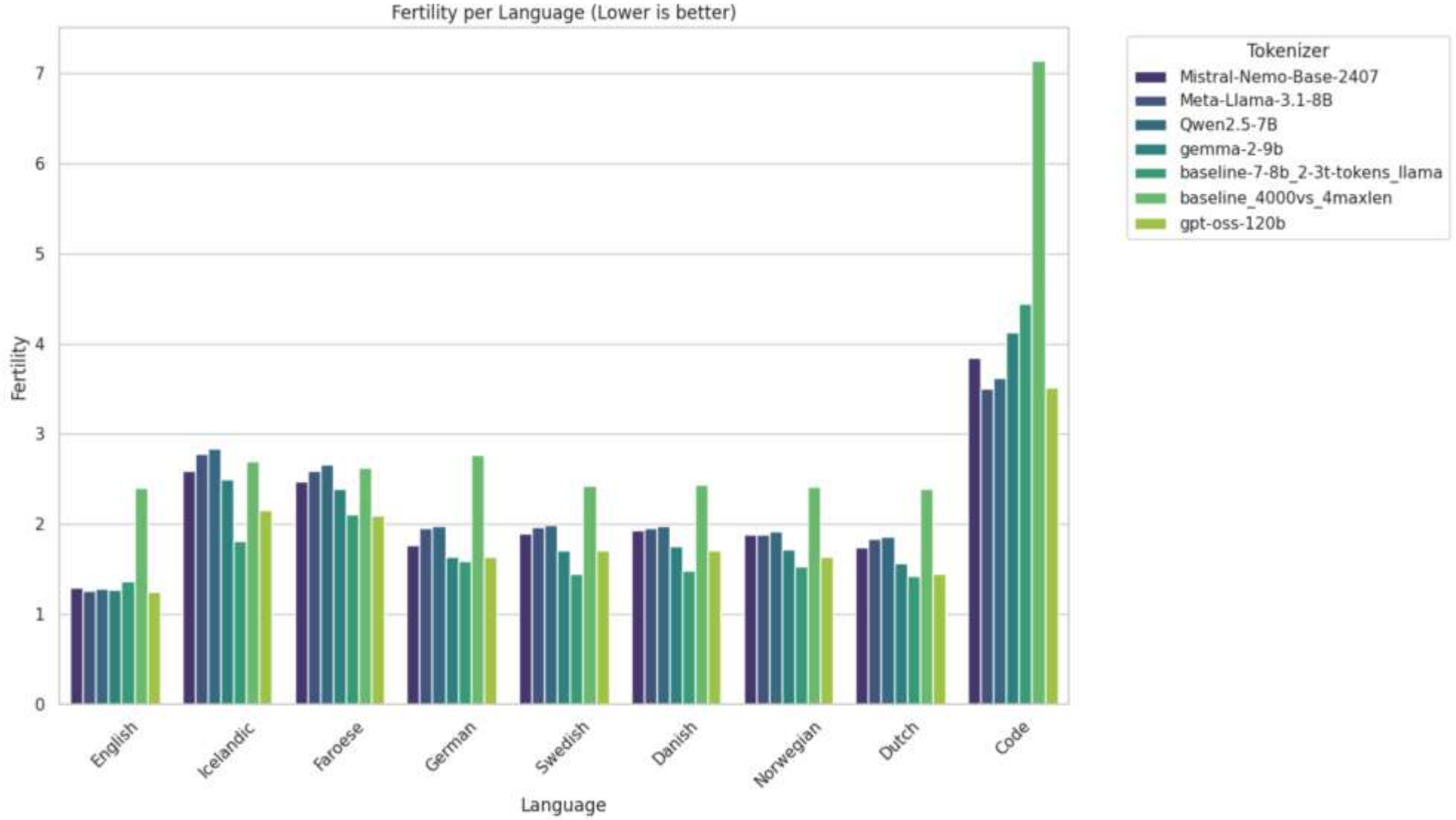
Funded by the European Union

Tokenisation Comparison for Germanic Languages



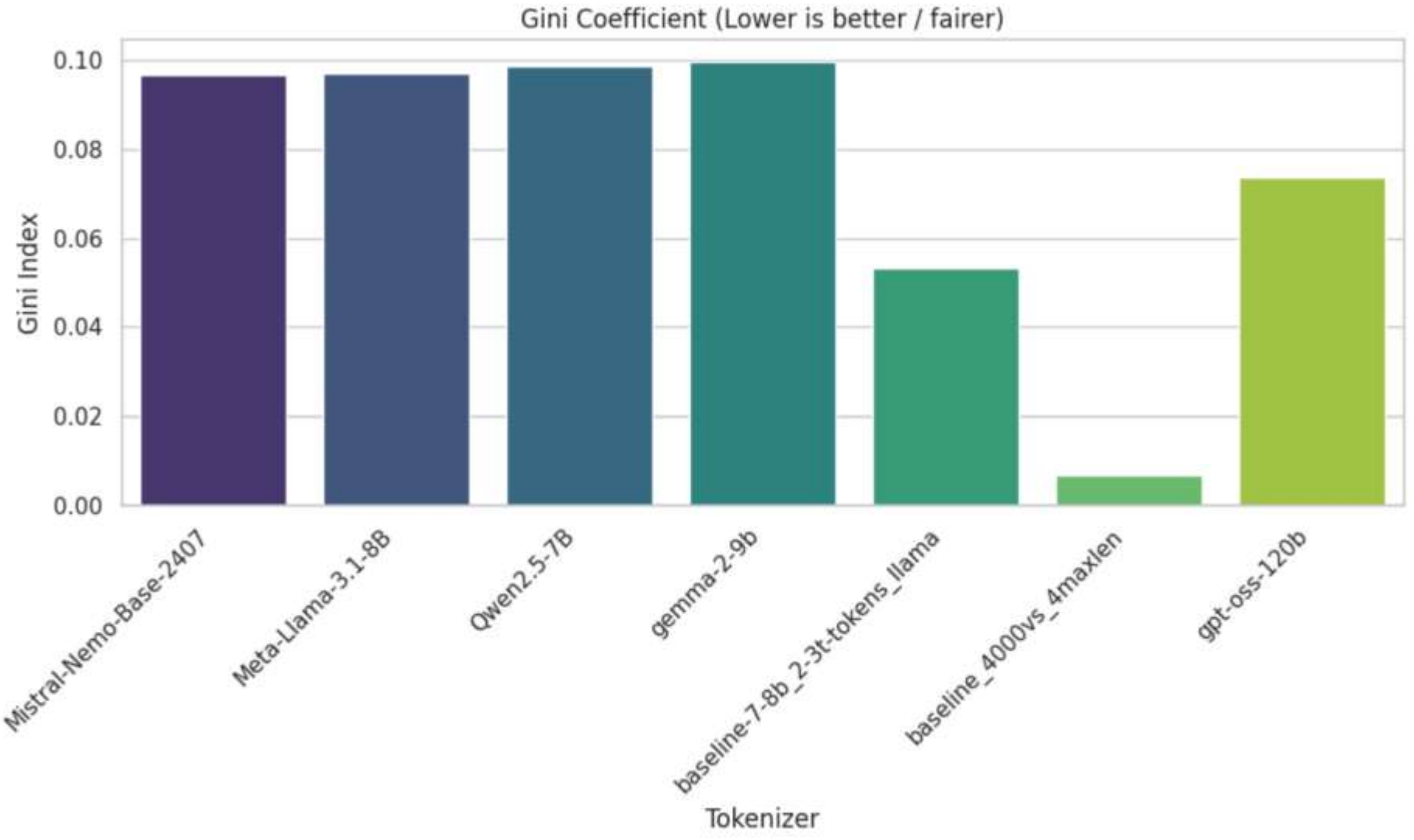
Funded by the European Union

Tokenisation Comparison for Germanic Languages



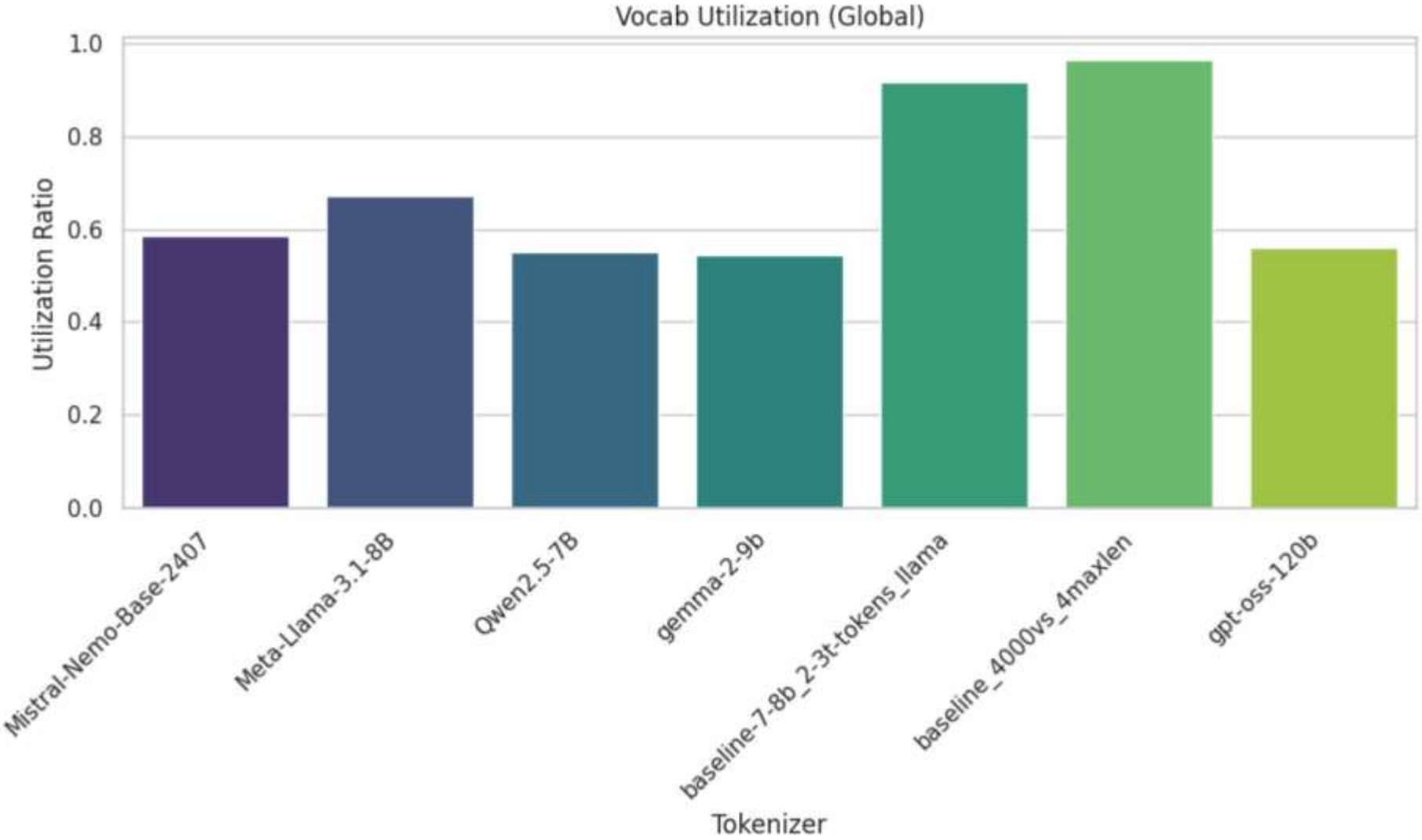
Funded by the European Union

Tokenisation Comparison for Germanic Languages



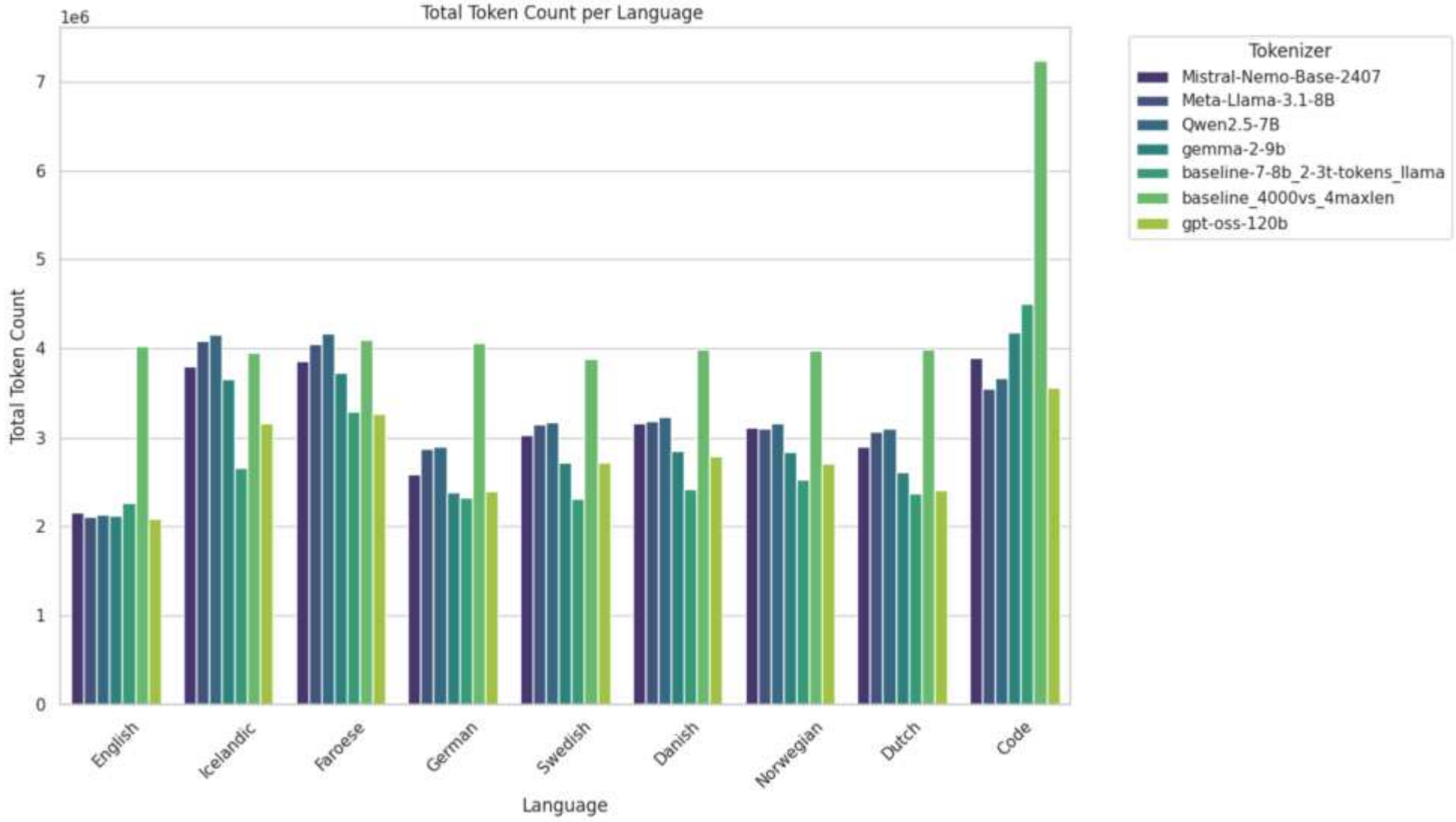
Funded by the European Union

Tokenisation Comparison for Germanic Languages



Funded by the European Union

Tokenisation Comparison for Germanic Languages

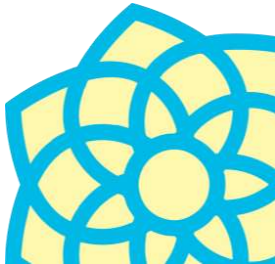


Funded by the European Union

Tokenisation Comparison for Germanic Languages

Use Case	Recommended Tokenizer	Rationale
English-focused	GPT-OSS-120b	Best English compression (5.03), lowest English fertility (1.25)
Nordic languages	TrustLLMeu baseline	Best compression/fertility for Icelandic, Swedish, Danish, Norwegian
Multilingual fairness	TrustLLMeu baseline	Lowest practical Gini (0.053) with good efficiency
General multilingual	GPT-OSS-120b	Good balance of English excellence and Nordic support (Gini 0.074)
Code	Meta-Llama-3.1-8B	Lowest code token count, good code fertility

Funded by the European Union



Tokenisation – Not only Text

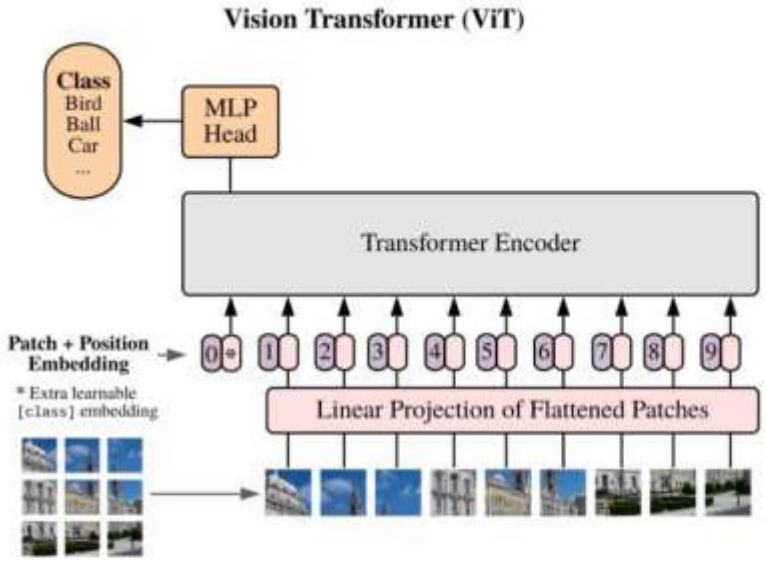
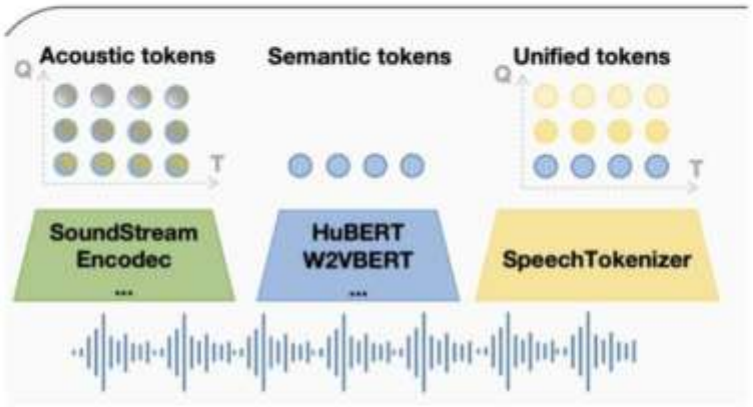
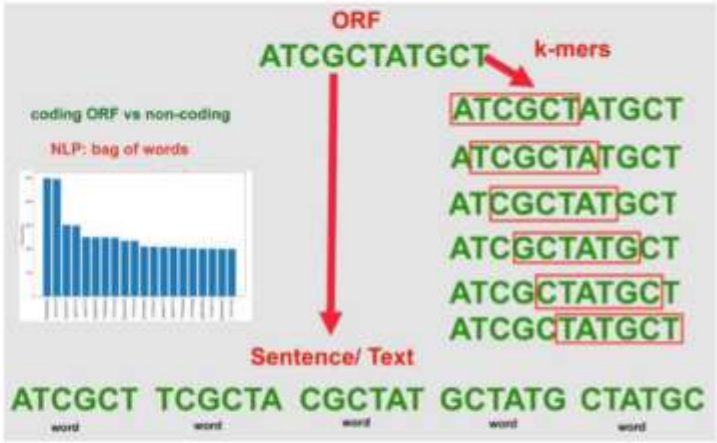


Image token



Speech token



genes （基因）

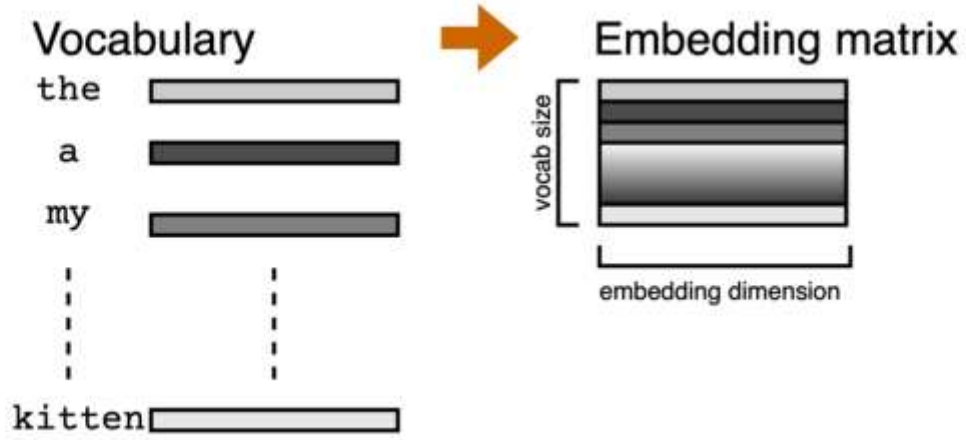
Alexey Dosovitskiy. et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://arxiv.org/abs/2010.11929
Xin zhang et.al. SpeechTokenizer: Unified Speech Tokenizer for Speech Language Models. https://0nutation.github.io/SpeechTokenizer.github.io/

Funded by the European Union

Turning Discrete Tokens into Continuous Vectors

Neural networks cannot operate on discrete tokens.

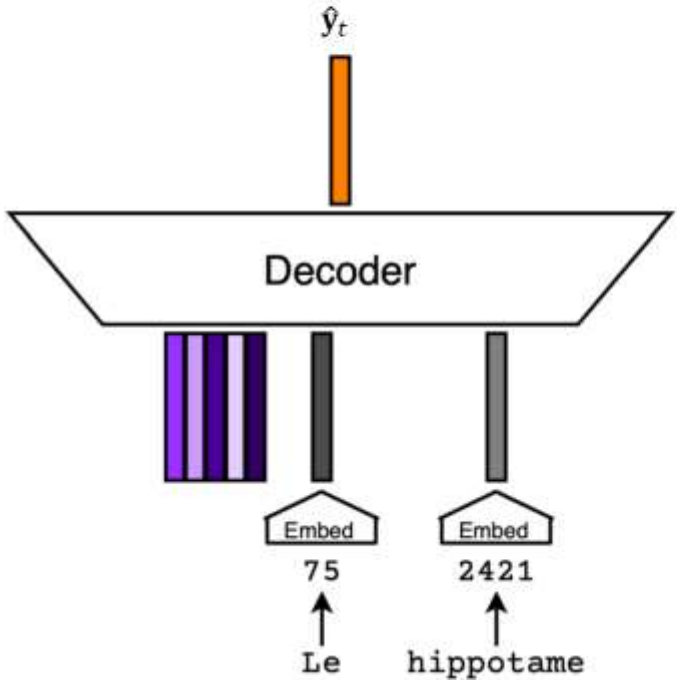
Instead, we build an **embedding matrix** which associates each token in the vocabulary with a vector embedding.

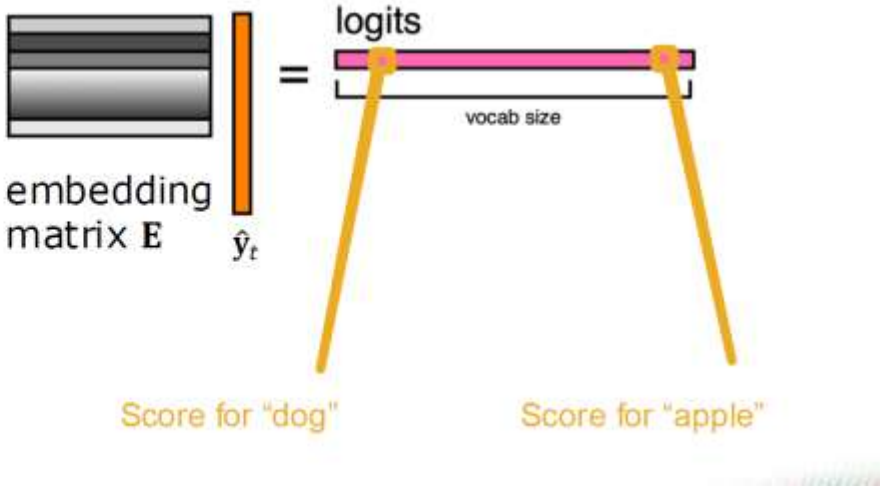


Funded by the European Union

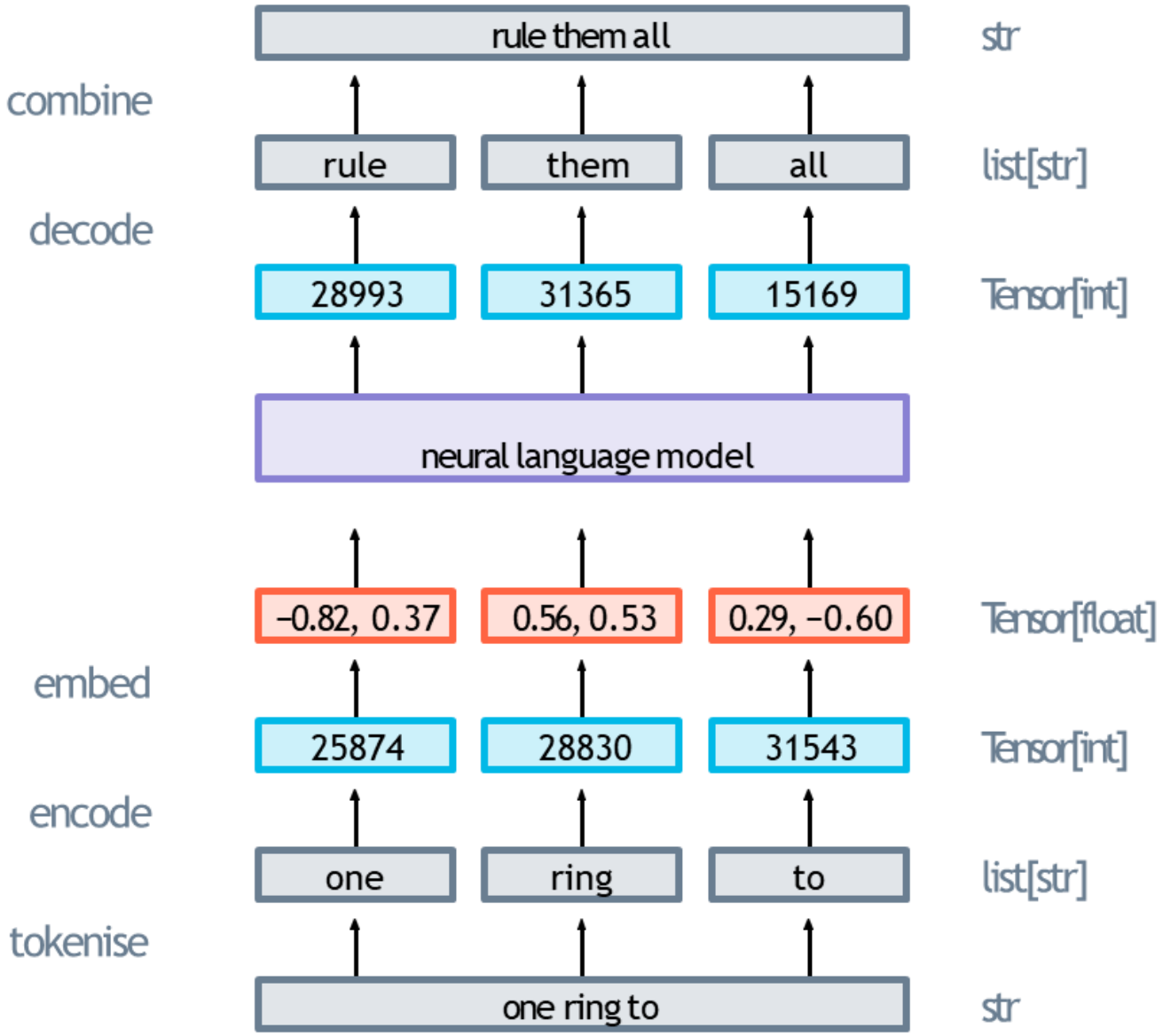
Decoder Inputs and Outputs

We multiply the predicted embedding $\hat{\mathbf{y}}_t$ by our vocabulary embedding matrix to get a score for each vocabulary word. These scores are referred to as **logits**.





Funded by the European Union



Funded by the European Union

Data Processing Pipeline

Funded by
the European Union

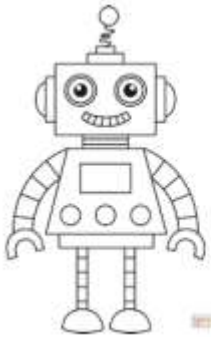
Data for LLM pretraining

- Training modern LLMs demands vast amounts of data. This data is often sourced from the Internet.
- While abundant, Internet data is unstructured, noisy, and biased, making it an imperfect representation of language.
- Internet text data requires extensive postprocessing and quality filtering to enhance relevance and diversity.

Funded by the European Union

Data and Large Language Model Training

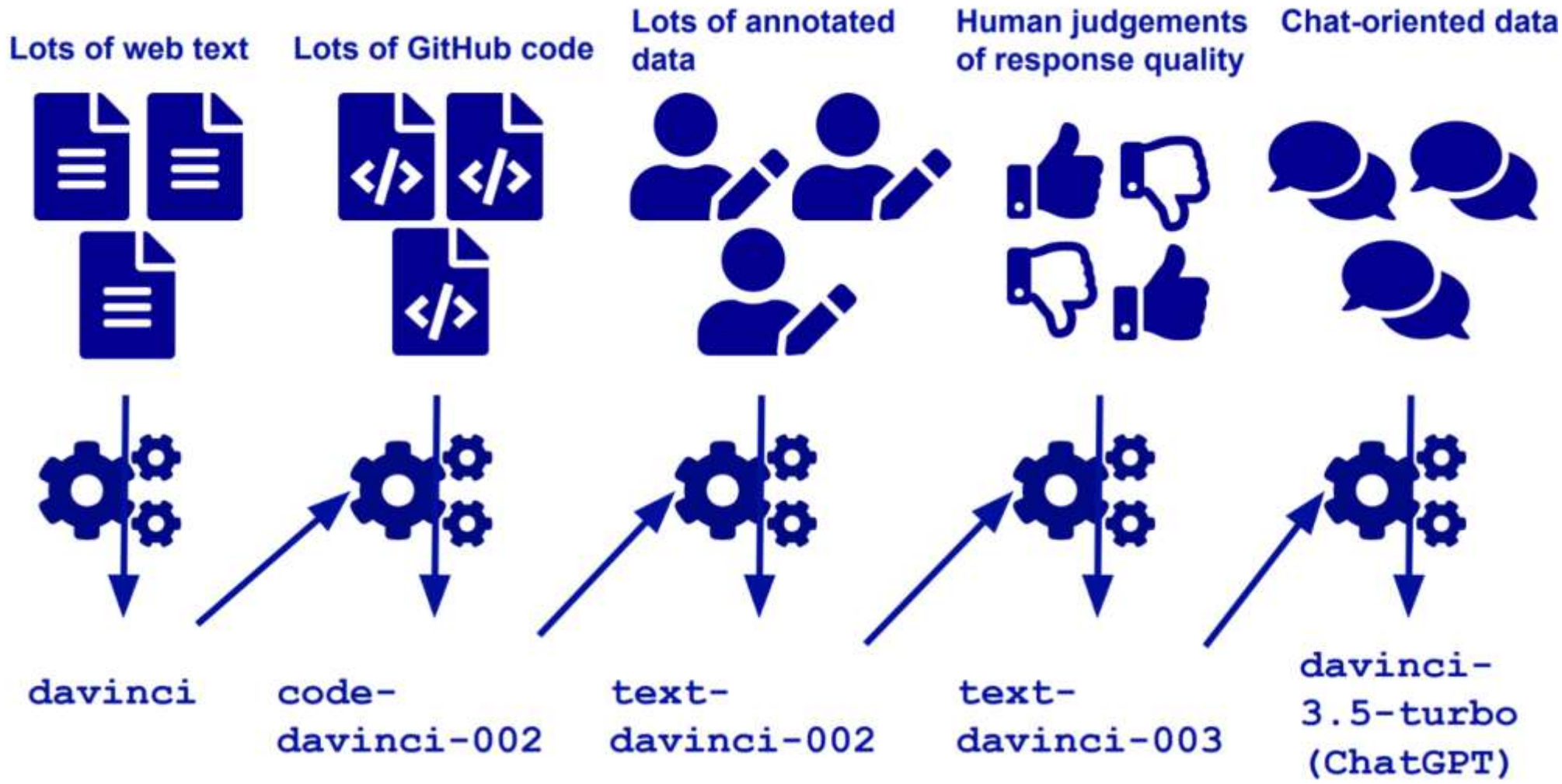
Recognize Words → TextBook Reading → Doing Exercises → Teachers' feedback



Tokenizer Training → Self-supervised Pre-training → Instruction Finetuning → Reinforcement Learning from Human Feedback

Funded by the European Union

Data and Large Language Model Training



Funded by the European Union

Data and Large Language Model Training

Table 1.1 The pretraining dataset of the popular GPT-3 LLM

Dataset name	Dataset description	Number of tokens	Proportion in training data
CommonCrawl (filtered)	Web crawl data	410 billion	60%
WebText2	Web crawl data	19 billion	22%
Books1	Internet-based book corpus	12 billion	8%
Books2	Internet-based book corpus	55 billion	8%
Wikipedia	High-quality text	3 billion	3%

Funded by the European Union

Pre-training Data Reality

- **Web data is plentiful, but can be challenging to work with.**
  - Copyright and usage constraints can get extremely complicated
  - Data is noisy, dirty, and biased
  - Data is contaminated with auto-generated text
    - Not just from LLM usage, but also tons of templated text.

Funded by the European Union

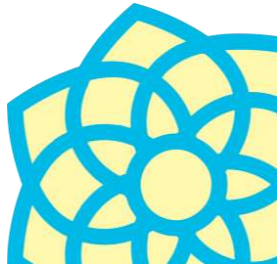
The Web Data Pipeline

1. Content is posted to the web.

2. Webcrawlers identify and download a portion of this content.

3. The data is filtered and cleaned.

Funded by the European Union



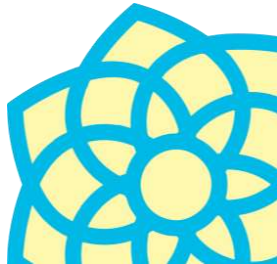
The Web Data Pipeline

1.Content is posted to the web.

2.Webcrawlers identify and download a portion of this content.

3.The data is filtered and cleaned.

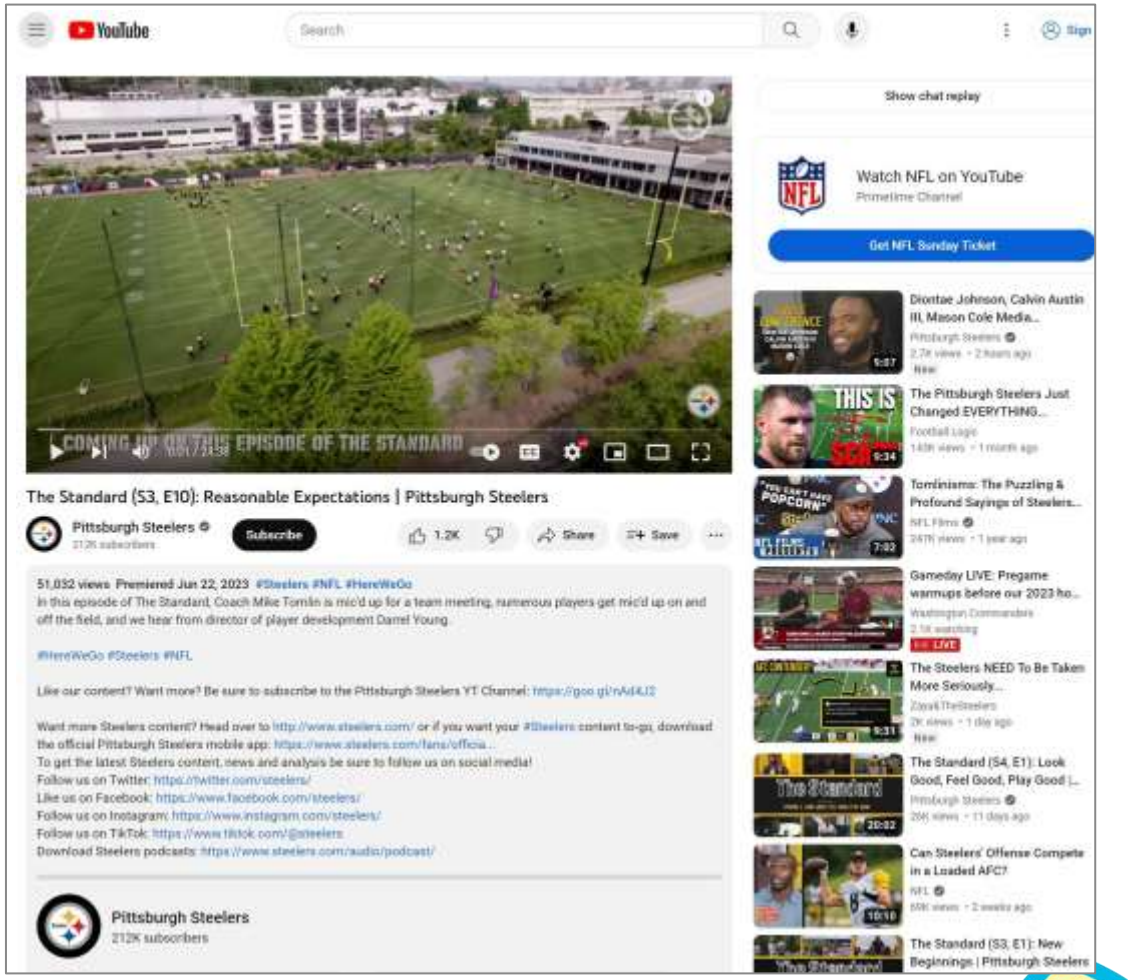
Funded by the European Union



1. Content is posted to the internet.

Challenges

- Biases in what's available:
  - Recency bias
  - Demographic biases
  - Language biases
- Web is much more dynamic than static HTML pages
  - CSS, JavaScript, interactivity, etc.
  - Responsive design
  - Many HTML pages involves 20+ secondary URLs, iframes, etc.
- What counts as content?
  - Ads, recommendation, navigation, etc.
  - Multimedia: images, videos, tables, etc.
  - Spam



1. Content is posted to the internet.

Content extraction from webpages is a well-studied problem in industry.

- Can be very engineering and resource heavy to do well.
- Existing toolkits is a strategic advantage of some proprietary LLMs

Figure 8: A Content Extraction Pipeline from Bing Used for ClueWeb22

Funded by the European Union

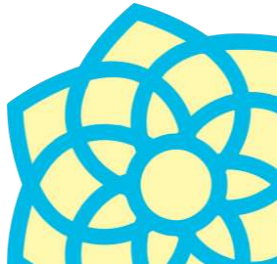
The Web Data Pipeline

1. Content is posted to the web.

2. **Webcrawlers identify and download a portion of this content.**

3. The data is filtered and cleaned.

Funded by the European Union



2. Webcrawlers identify and download a portion of this content.

General Idea

1. Start with a set of seed websites
2. Explore outward by following all hyperlinks on the webpage.
3. Systematically download each webpage and extract the raw text.

2. Webcrawlers identify and download a portion of this content.

General Idea

1. Start with a set of seed websites
2. Explore outward by following all hyperlinks on the webpage.
3. Systematically download each webpage and extract the raw text.

Challenges

- How to harvest a large number of seed URLs efficiently
- How to select “high quality” URLs and skip over “bad” URLs
  - Some cases are clear cut: spammy, unsafe, etc.
  - Some are hard to detect or up to debate: certain biases.
- How to keep the crawl up-to-date
  - Given a fixed compute budget each month, is it better to crawl new webpages, or recrawl old ones that might’ve changed?

Funded by the European Union

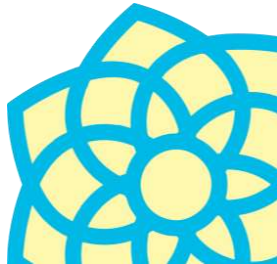
The Web Data Pipeline

1. Content is posted to the web.

2. Webcrawlers identify and download a portion of this content.

3. The data is filtered and cleaned.

Funded by the European Union



3. The data is filtered and cleaned.

Remove noisy, spammy, templated, and and fragmented texts

- These portions lack the content needed to meet pretraining goals.

Select higher quality texts from a massive candidate pool

- Given a limited pretraining compute budget, we'd like pretrain on better texts

Avoid toxic and biased content

- NSFW content
- Texts with strong biases

Funded by the European Union

3. The data is filtered and cleaned.

Methods to identify high-quality content:

- Rule-based heuristics for quality
  - You'll be implementing a bunch of these in the homework
- Proximity to known high-quality indicators
  - E.g. a website that was highly upvoted on Reddit, or a website that was included as a reference on Wikipedia
- Classifiers
  - Trained quality and toxicity classifiers are common.
  - E.g. train a classifier with Wikipeida as the positive examples and random web pages as the negative examples

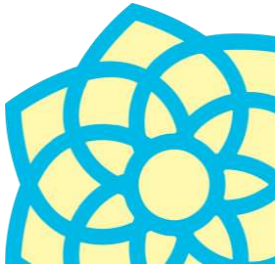
Funded by the European Union

3. The data is filtered and cleaned.

Example: Rule-based filtering in C4 (pre-training dataset for T5 model)

1. Start from Common Crawl's official extracted texts from HTML
2. Only keep text lines ended with a terminal punctuation mark
3. Discard pages with fewer than 5 sentences
4. Only keep lines with at least 3 words
5. Remove any line with the word "Javascript"
6. Remove any page
    1. with any words in a toxic word dictionary
    2. with the phrase "lorem ipsum"
    3. With "{"
7. De-dup at three-sentence span level

Funded by the European Union



3. The data is filtered and cleaned.

Challenges

- At what granularity should filtering be performed?
  - Word-level, sentence-level, paragraph-level, document level?
- What constitutes “high quality” or “non-toxic”?
- Are our filters/classifiers multilingual? Even within English, do they treat all groups equally?
- It is very expensive to ablate pre-training dataset decisions.

Funded by the European Union

Case Study: CommonCrawl

Common Crawl: commoncrawl.org

- Non-profit organization provides open access to large scale web crawls
- Petabytes of web pages available
- Monthly crawls and dumps
  - Re-crawled web pages and fresh dumps (bi)monthly
  - Recent dumps are ~3 billion pages
  - Date back to past 10 years
    - “220 billion web pages (HTML) captured 2008 – 2021”

Funded by the European Union

Case Study: CommonCrawl

Web exploration approach

- Start from a set of seed URLs: popular, high-quality, and trustworthy websites
  - Gov, Edu, etc.
  - Top web domains
- Traverse the web to obtain a candidate set of URLs
  - Around 500 billion links discovered per month crawl
  - About 25+ billion unique ones
- Prioritizes a subset (~3 billion) of URLs to crawl and include in the dump
  - Does this to make the best use of crawling budget.

Funded by the European Union

Case Study: CommonCrawl

Which URLs to crawl?

- 2008-2012: CC's in-house Page Rank
- 2012-2015: Added ranking and metadata of 22 billion pages donated from web search engine blekko
- 2016-2018: Occasional seed URL donations, ~400 million URLs
- 2016: Alexa and Common Search Rankings
- 2017-Now: CC's in-house web graph based rankings (page rank and centrality)
  - Importance score calculated based on past three-month dumps
  - Steer the crawler for next three month
  - Capped # of URLs per domain

Funded by the European Union

Case Study: CommonCrawl

Pros

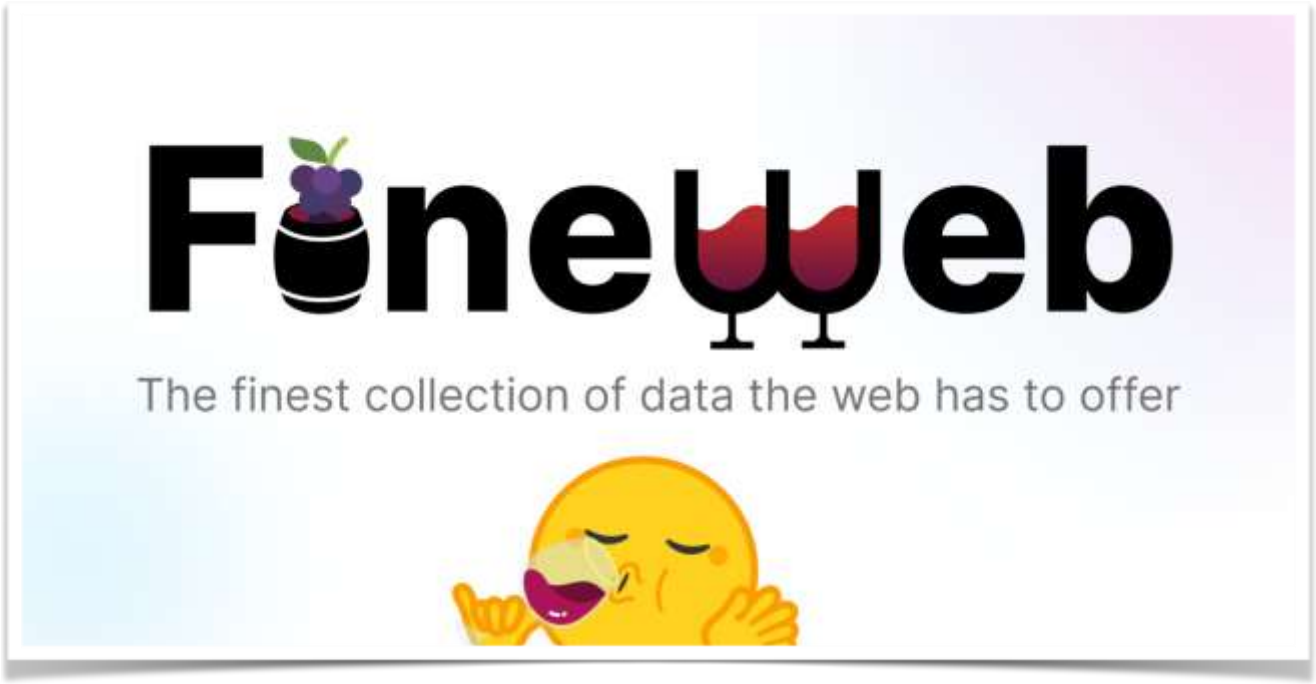
- The one and the only public web crawl at this scale
  - Still far away from commercial search engines, but the closest we have
- 10 years of crawl dumps have enabled various data subsamples
- Accumulation of low resource languages
  - One can combine low resources languages from years of dumps to pair with English texts from one month

Cons

Each crawl is ~3billion documents, still limited coverage of the massive web

- Crawls every month restart from the seed URLs
  - Crawls never grow above ~3B documents
- URL distributions skewed by crawling prioritization, per domain URL cap, and physical location of the crawling machines (and people)

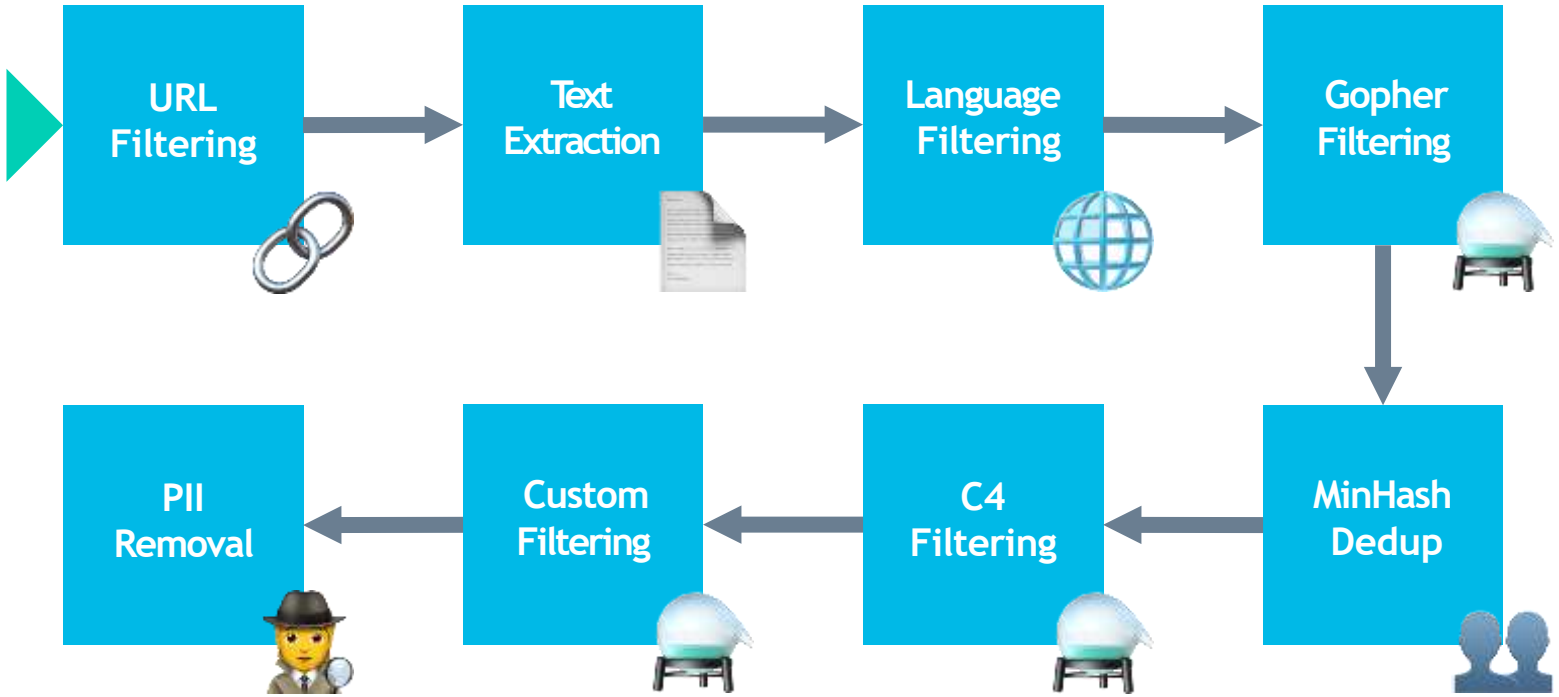
Funded by the European Union



”15 trillion tokens of the finest data the web has to offer”

Source

The FineWeb pipeline



Penedo et al. (2024)

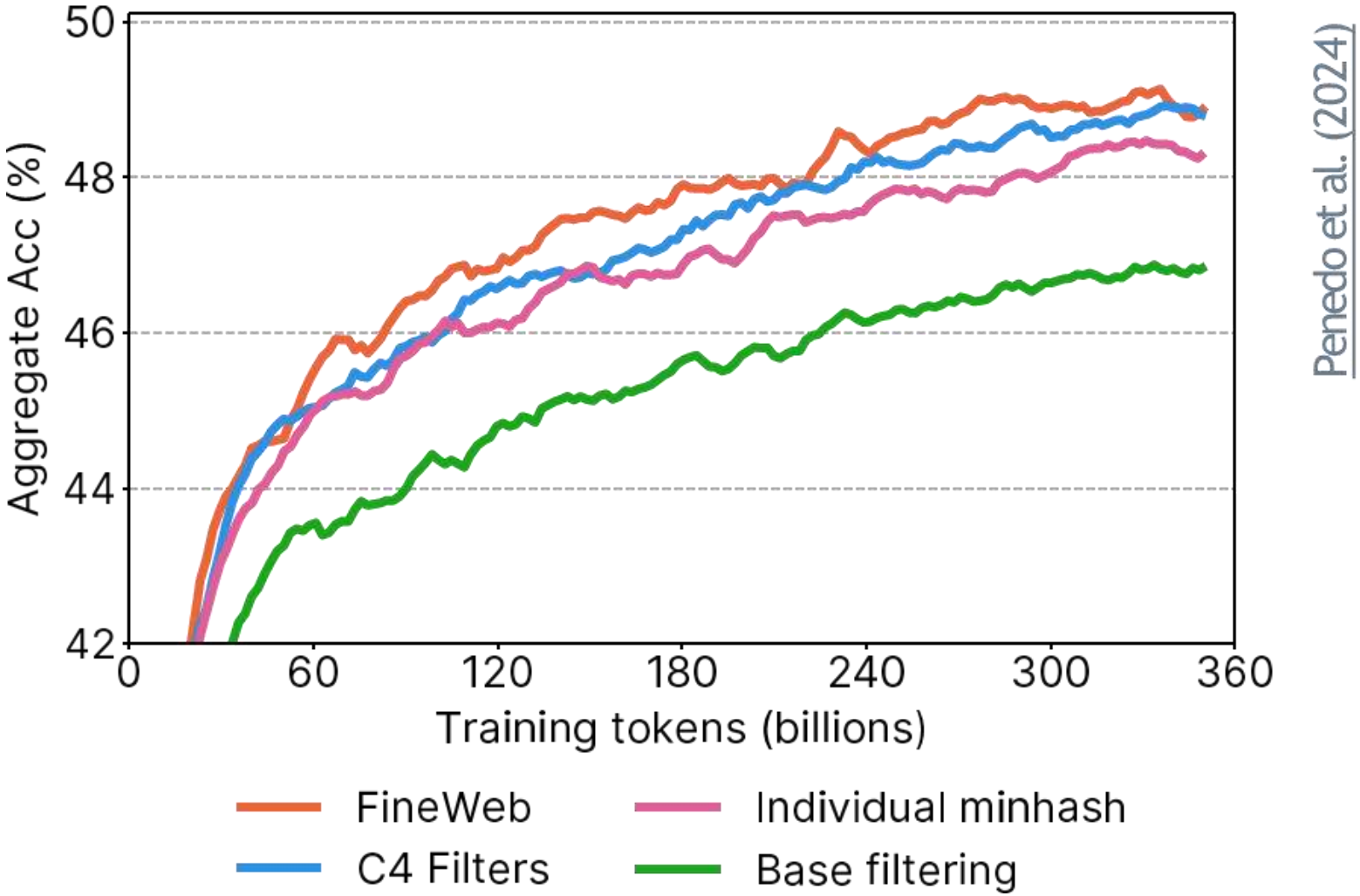
Funded by the European Union

Basic filtering

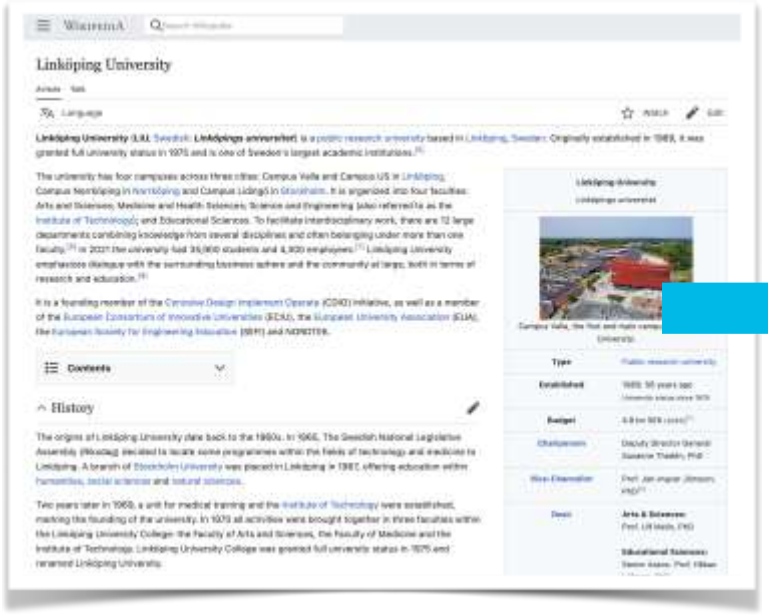
- URL filtering using blocklists
  Examples: adult content, malware, phishing sites
- Language filtering
  Uses fastText; keep only English text with a score $\geq 65\%$
- Heuristic filtering
  Filter for length, symbol-to-word ratio, common/uncommon words, etc.

Funded by the European Union

Impact of filtering



Text extraction



Linköping University (LiU; Swedish: Linköpings universitet) is a public research university based in Linköping, Sweden. Originally established in 1969, it was granted full university status in 1975 and is one of Sweden's largest academic institutions.[5]

Linköpings universitet | |
Type	Public research university

FineWeb uses Trafilatura.

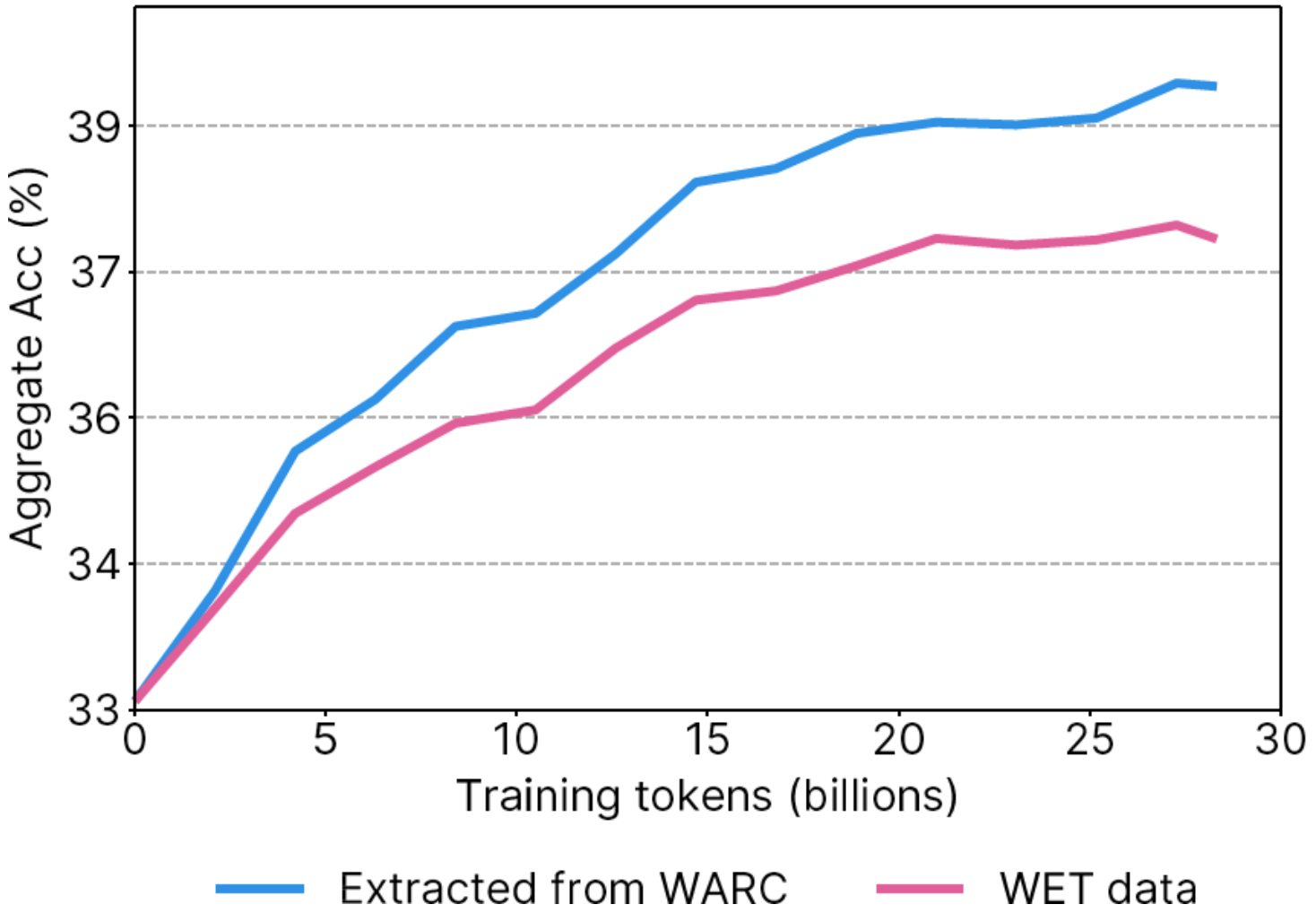
Funded by the European Union

Data Formats (WARC, WAT, WET)

- WARC (Web ARChive), industry standard for web archiving.
  - Can store multiple resources, similar to ARC, but with more capabilities. This includes the ability to store request and response headers, additional metadata, and new record types like resource revisit, metadata, and conversion.
- WAT (Web ARChive Timestamp).
  - Focus on the metadata associated with the crawled web pages. Contain parsed data from the HTTP response headers, links extracted from HTML pages, and other metadata. This can include information like server response codes, content types, languages, and more.
- WET (Web Extracted Text).
  - Contain extracted plain text from web content, the body text of web pages, extracted from the HTML and excluding any HTML code, images, or other media. This makes them useful for text analysis and natural language processing (NLP) tasks.

Funded by the European Union

Relevance of text extraction



Penedo et al. (2024)

Funded by the European Union

Deduplication

- Large-scale web datasets contain significant amounts of duplicate content, which can lead to overfitting.
- Deduplication leads to a more diverse dataset and reduces computational cost.
- Deduplicating massive datasets requires efficient similarity detection techniques or other fuzzy approaches.
  embedding-based similarity or MinHash

Funded by the European Union

De-duplication

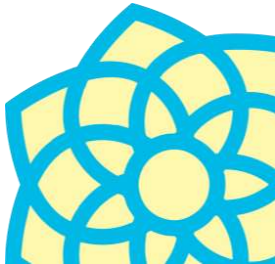
Dataset	Example	Near-Duplicate Example
Wiki-40B	\n_START_ARTICLE_\nHum Award for Most Impactful Character \n_START_SECTION_\nWinners and nominees\n_START_PARAGRAPH_\nIn the list below, winners are listed first in the colored row, followed by the other nominees. [...]	\n_START_ARTICLE_\nHum Award for Best Actor in a Negative Role \n_START_SECTION_\nWinners and nominees\n_START_PARAGRAPH_\nIn the list below, winners are listed first in the colored row, followed by the other nominees. [...]
LM1B	I left for California in 1979 and tracked Cleveland 's changes on trips back to visit my sisters .	I left for California in 1979 , and tracked Cleveland 's changes on trips back to visit my sisters .
C4	Affordable and convenient holiday flights take off from your departure country, "Canada". From May 2019 to October 2019, Condor flights to your dream destination will be roughly 6 a week! Book your Halifax (YHZ) - Basel (BSL) flight now, and look forward to your "Switzerland" destination!	Affordable and convenient holiday flights take off from your departure country, "USA". From April 2019 to October 2019, Condor flights to your dream destination will be roughly 7 a week! Book your Maui Kahului (OGG) - Dubrovnik (DBV) flight now, and look forward to your "Croatia" destination!

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C. and Carlini, N., 2021. Deduplicating training data makes language models better. arXiv preprint arXiv:2107.06499.

Funded by the European Union

Name	Based on	Release year	Number of tokens
C4	Common Crawl	2019	156B
WebText	Own Crawl (OpenAI)	2019	300B
CC-100	Common Crawl	2020	532B
MassiveText	Own Crawl (Google)	2022	2.3T
OSCAR	Common Crawl	2023	523B
RedPajama	Common Crawl	2023	30.4T
RefinedWeb	Common Crawl	2023	500B
Dolma	Common Crawl	2024	3T
FineWeb	Common Crawl	2024	15T

Funded by the European Union



FineWeb and FineWeb-EDU

- Design choices were validated through training data ablation studies and evaluated on downstream task benchmarks.
- The authors released a 1.3T-token filtered subset of FineWeb, focusing on high-quality educational web pages.
- FineWeb-EDU was built using an educational quality classifier, trained on labels generated by Llama (1.71B).
  linear regression on top of an embedding model

- Penedo et al. (2024

Funded by the European Union

Below is an extract from a web page. Evaluate whether the page has a high educational value and could be useful in an educational setting for teaching from primary school to grade school levels using the additive 5-point scoring system described below. Points are accumulated based on the satisfaction of each criterion:

- Add 1 point if the extract provides some basic information relevant to educational topics, even if it includes some irrelevant or non-academic content like advertisements and promotional material.

- Add another point if the extract addresses certain elements pertinent to education but does not align closely with educational standards. It might mix educational content with non-educational material, offering a superficial overview of potentially useful topics, or presenting information in a disorganized manner and incoherent writing style.

- Award a third point if the extract is appropriate for educational use and introduces key concepts relevant to school curricula. It is coherent though it may not be comprehensive or could include some extraneous information. It may resemble an introductory section of a textbook or a basic tutorial that is suitable for learning but has notable limitations like treating concepts that are too complex for grade school students.

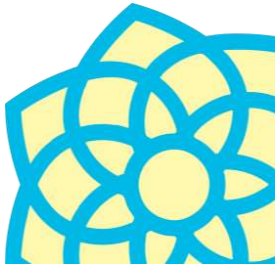
- Grant a fourth point if the extract highly relevant and beneficial for educational purposes for a level not higher than grade school, exhibiting a clear and consistent writing style. It could be similar to a chapter from a textbook or a tutorial, offering substantial educational content, including exercises and solutions, with minimal irrelevant information, and the concepts aren't too advanced for grade school students. The content is coherent, focused, and valuable for structured learning.

- Bestow a fifth point if the extract is outstanding in its educational value, perfectly suited for teaching either at primary school or grade school. It follows detailed reasoning, the writing style is easy to follow and offers profound and thorough insights into the subject matter, devoid of any non-educational or complex content.

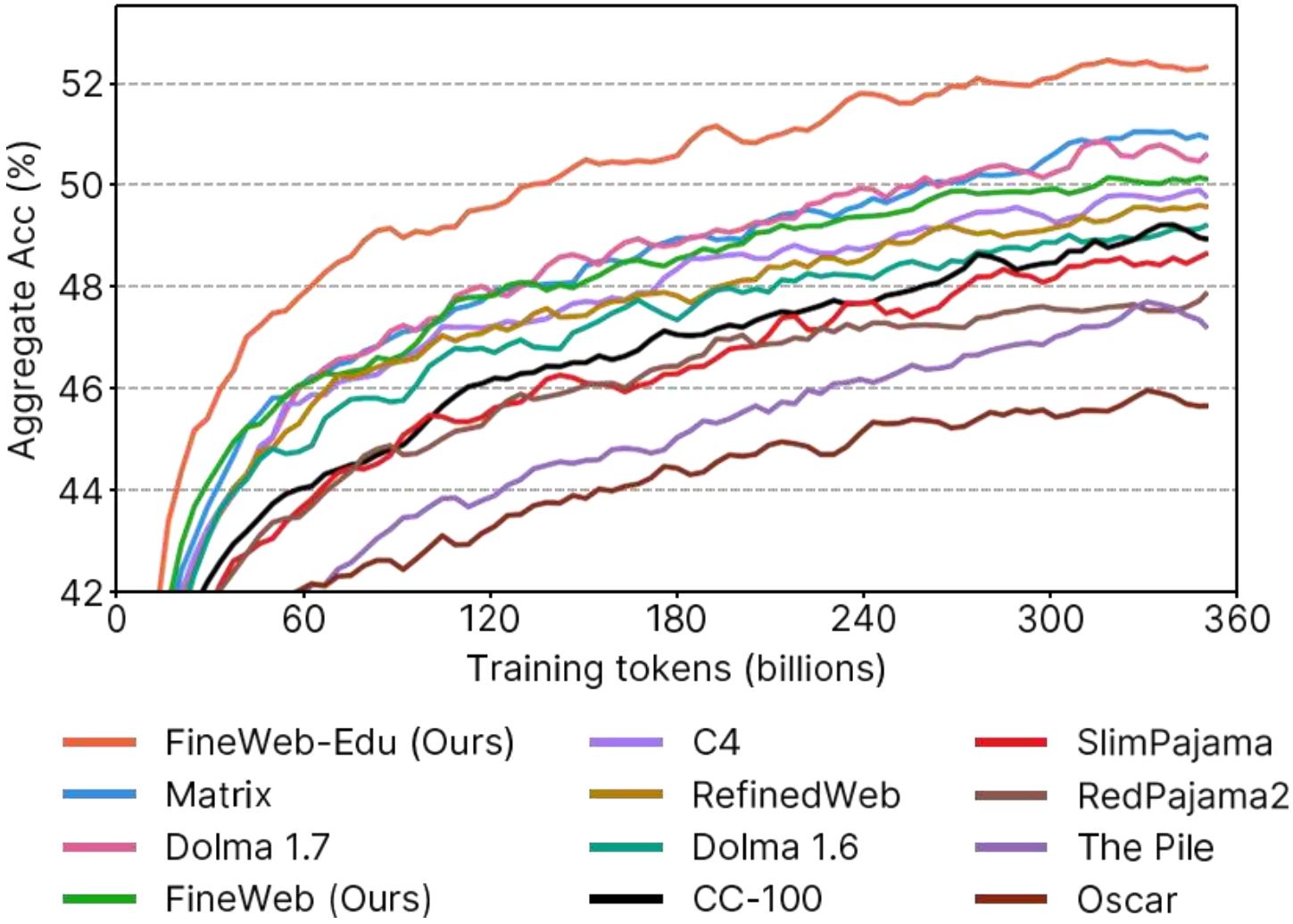
The extract: <EXAMPLE>. After examining the extract:

- Briefly justify your total score, up to 100 words.
- Conclude with the score using the format: "Educational score: <total points>"

Funded by the European Union



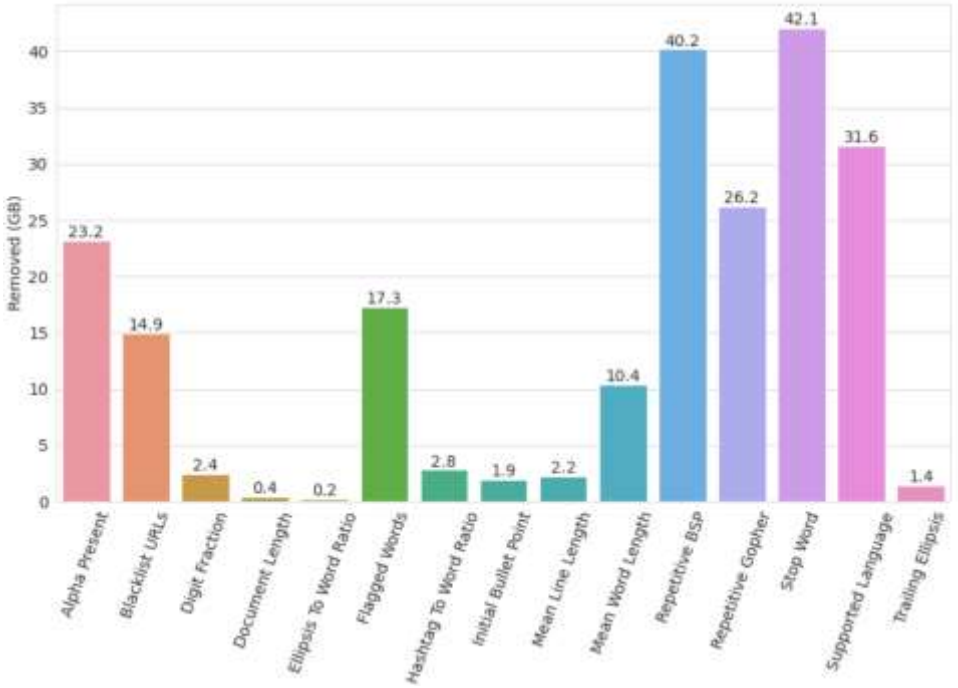
Impact of high-quality data

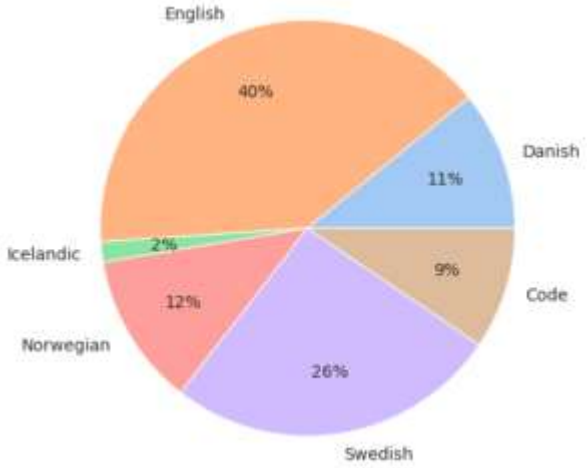


Penedo et al. (2024)

Funded by the European Union

The Nordic Pile: A 1.2TB Nordic Dataset





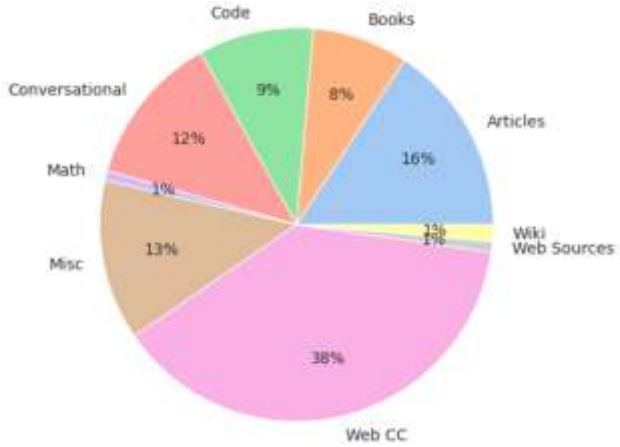


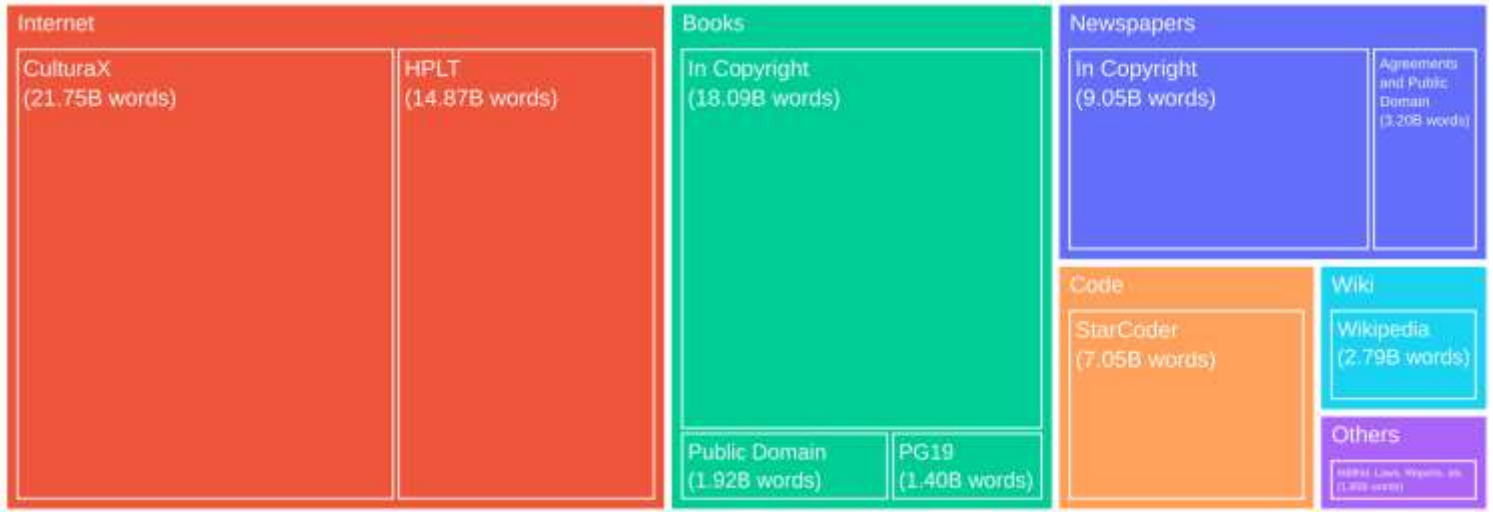
Table 2: Data sizes for each language and category.

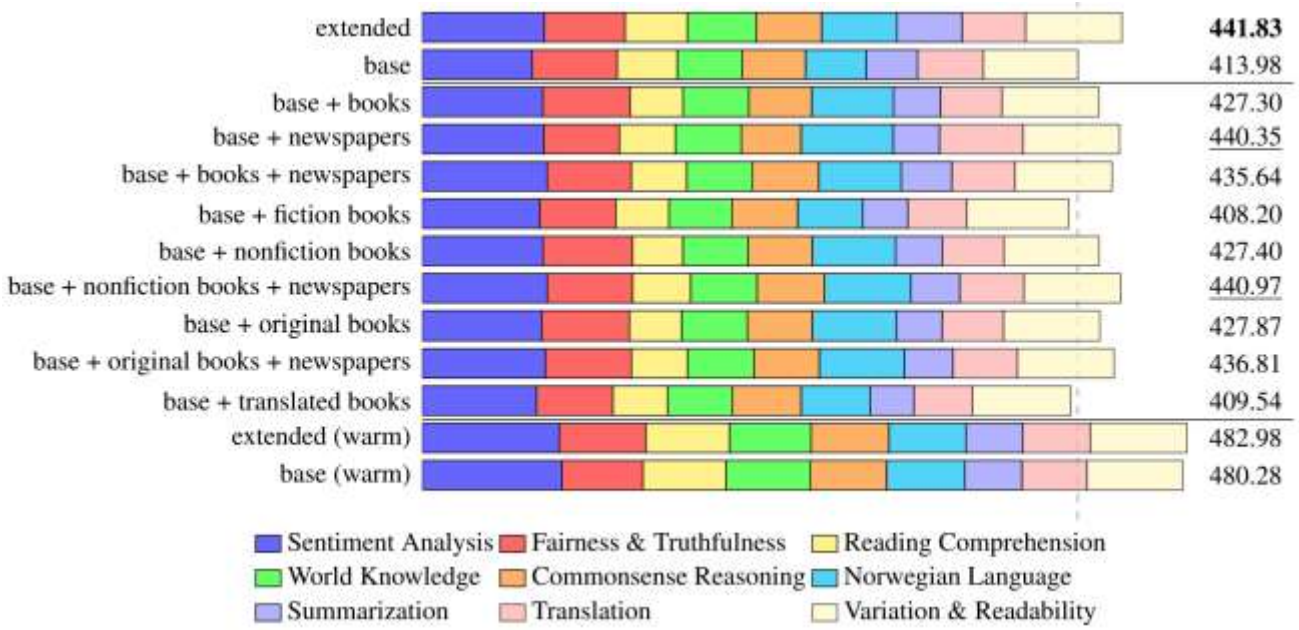
	Danish	English	Icelandic	Norwegian	Swedish	Other	Code	Total
Articles	0.19 GB	173.52 GB	0 GB	0.01 GB	16.49 GB	0 GB		190.21 GB
Books	0.06 GB	94.14 GB	0 GB	0.04 GB	1.15 GB	0 GB		95.39 GB
Conversational	2.84 GB	81.67 GB	0.07 GB	0.57 GB	65.61 GB	0.01 GB		150.77 GB
Math	0.01 GB	4.98 GB	0 GB	0.01 GB	4.58 GB	0.19 GB		9.77 GB
Miscellaneous	13.85 GB	56.31 GB	10.26 GB	48.48 GB	28.85 GB	1.8 GB		159.55 GB
Web CC	111.33 GB	60.36 GB	8.79 GB	90 GB	188.94 GB	2.05 GB		461.47 GB
Web Sources	1.85 GB	0.61 GB	0 GB	0.03 GB	7.83 GB	0 GB		10.32 GB
Wikipedia	0.38 GB	14.77 GB	0.05 GB	0.48 GB	1.03 GB	0 GB		16.71 GB
Code							114.5 GB	114.5 GB
Total	130.51 GB	486.36 GB	19.17 GB	139.62 GB	314.48 GB	4.05 GB	114.5 GB	1208.69 GB

The Nordic Pile: A 1.2TB Nordic Dataset for Language Modeling. 2023. https://arxiv.org/abs/2303.17183

Funded by the European Union

The Impact of Copyrighted Material on LLMs: A Norwegian Perspective





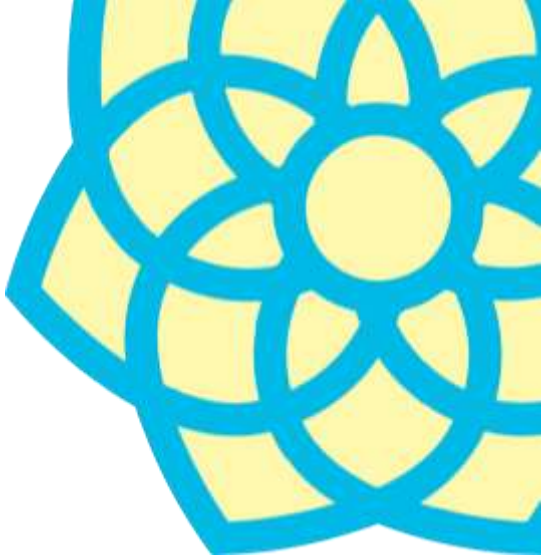
Dataset	Documents	Words
base	60,182,586	40,122,626,81
extended	125,285,547	82,149,281,266

Subset	Documents	Words
books	492,281	18,122,699,498
newspapers	46,764,024	9,001,803,515
books + newspapers	47,256,305	26,078,915,554
fiction books	117,319	5,287,109,366
nonfiction books	359,979	12,384,323,012
nonfiction books + newspapers	42,083,532	20,340,539,068
original books	392,887	13,352,261,605
original books + newspapers	47,156,911	22,354,065,120
translated books	96,258	4,695,814,506

TrustLLM

The Impact of Copyrighted Material on Large Language Models: A Norwegian Perspective. De la Rosa, etal. 2025. https://arxiv.org/abs/2412.09460v1

Funded by the European Union



TrustLLM Data Processing

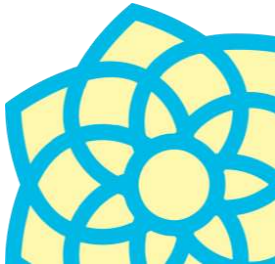
TrustLLM

Funded by
the European Union

Data Sources - JUDAC

Dataset	Size	Tokens (GPT-OSS)	Primary Language
tdm/fineweb-edu	738 GB	157.6 BT	English
tdm_not_checked/fineweb-edu	1.8 TB	384.3 BT	English
tdm/fineweb2	888 GB	225.6 BT	Mixed Germanic (55% deu, 20% nld, 9% nor, 8% swe, 8% dan)
tdm_not_checked/fineweb2	183 GB	44.9 BT	German
common_corpus	1.2 TB	252.8 BT	Mixed (85% eng, 13% deu, 2% other)
code	851 GB	309.1 BT	Code (358 languages)
hplt2/**HPLT3**	628 GB	158.8 BT	Mixed (69% deu, 13% nld, 10% swe, 4% dan, 4% nor)
curated	270 GB	60.9 BT	Mixed (83% eng, 6% deu, 5% swe, 4% isl, 2% other)
finemath	140 GB	29.9 BT	English
fao_last_minute	77 MB	26 MT	Faroese
Total	**~6.7 TB**	**~1.62 TT**	**8 Germanic + Code**

Funded by the European Union



Data Sources – MultiSynt

Dataset	Size	Tokens	Primary Language
finepdfs-summaries	318 GB	72.8 BT	Mixed (51% eng, 47% deu, 2% other Germanic)
MT-HPLT2c	502 GB	121.4 BT	Mixed (40% eng, 31% deu, 28% swe)
MT-Nemotron-CC (parallel)	1.3 TB	338.2 BT	Multi-Germanic (21% deu, 20% swe, 19% nor, 10% nld)
MT-Reasoning	89 GB	20.5 BT	Mixed (53% deu, 47% eng)
MT-Reasoning-Prompts	10 GB	2.6 BT	Mixed (34% deu, 33% nld, 32% dan)
Total (training data)	**~2.2 TB**	**~555 BT**	**Germanic + English**
Excluded	~6.6 TB	—	French, Spanish, Italian, Portuguese, Polish, Romanian, Hungarian, Czech, Finnish, Ukrainian (no compression ratios)

Funded by the European Union

Data Sources – Nemotron-Pretraining-Specialized-v1

Dataset	Size	Tokens	Content Type
Nemotron-Pretraining-RQA	172 GB	36.7 BT	English (Q&A)
Nemotron-Pretraining-STEM-SFT	91 GB	19.4 BT	English (STEM)
Nemotron-Pretraining-Math-Textbooks	29 GB	6.2 BT	English (Math)
Nemotron-Pretraining-InfiniByte-Reasoning	27 GB	5.8 BT	English (Reasoning)
Nemotron-Pretraining-Wiki-Rewrite	9.3 GB	2.0 BT	English (Wiki)
Nemotron-Pretraining-Scientific-Coding	524 MB	0.2 BT	Code
Total	**~329 GB**	**~70 BT**	**English + Code**

Funded by the European Union

Data Sources – Nemotron-CC-Math-v1

Dataset	Size	Tokens	Content Type
3	101 GB	21.6 BT	Math (quality level 3)
4plus_MIND	85 GB	18.1 BT	Math (quality 4+ MIND)
4plus	58 GB	12.4 BT	Math (quality level 4+)
Total	**244 GB**	**52 BT**	**English (Math)**

Funded by the European Union

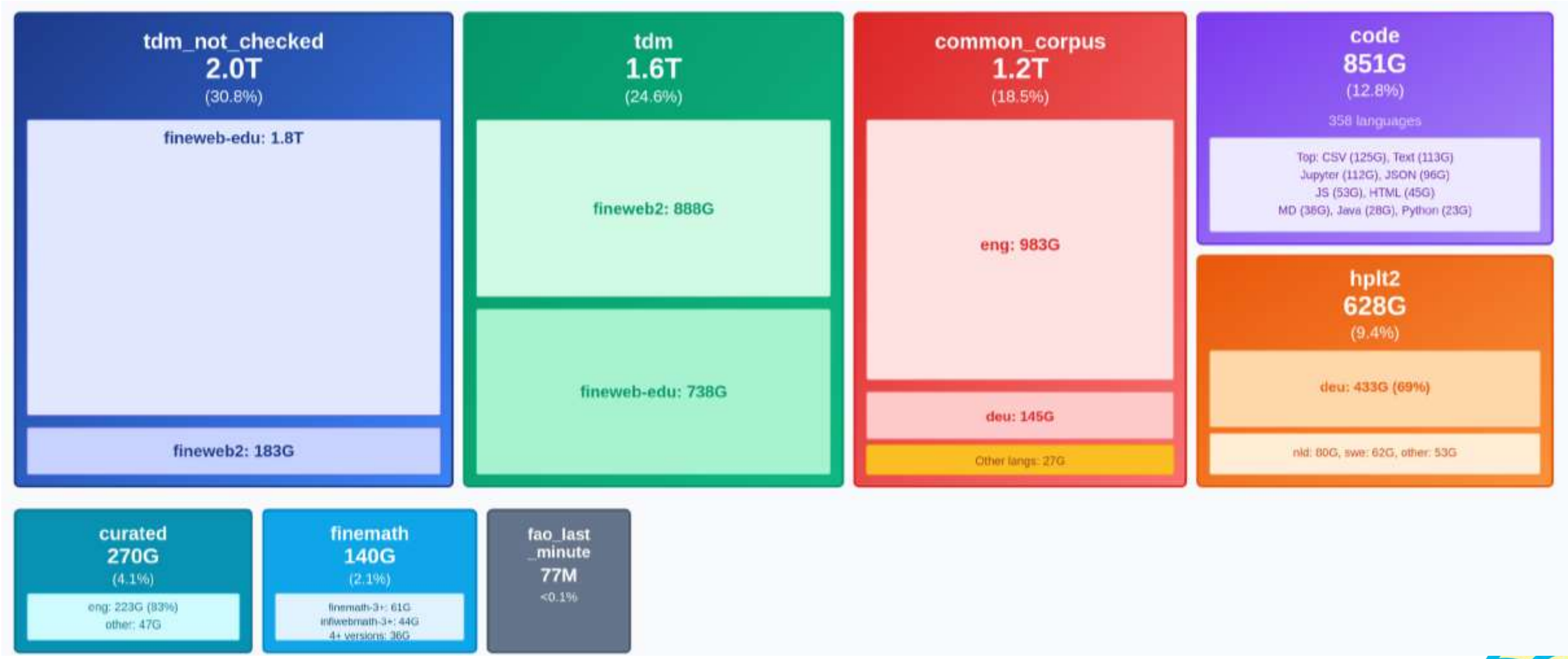
Data Sources – Dolma 3

- allenai/dolma3_mix-6T
- allenai/dolma3_dolmino_mix-100B-1025
- allenai/dolma3_longmino_mix-50B-1025

- Total tokens
  - **~8.384TT** = data_v2(1.62TT) + multisynt (0.555TT) + nemotron-pretrained-specialized-v1(0.007TT) + nemotron-cc-math-v1(0.052TT) + mix(6TT) + dolmino_mix(0.1TT) + longmino_mix(0.05TT)

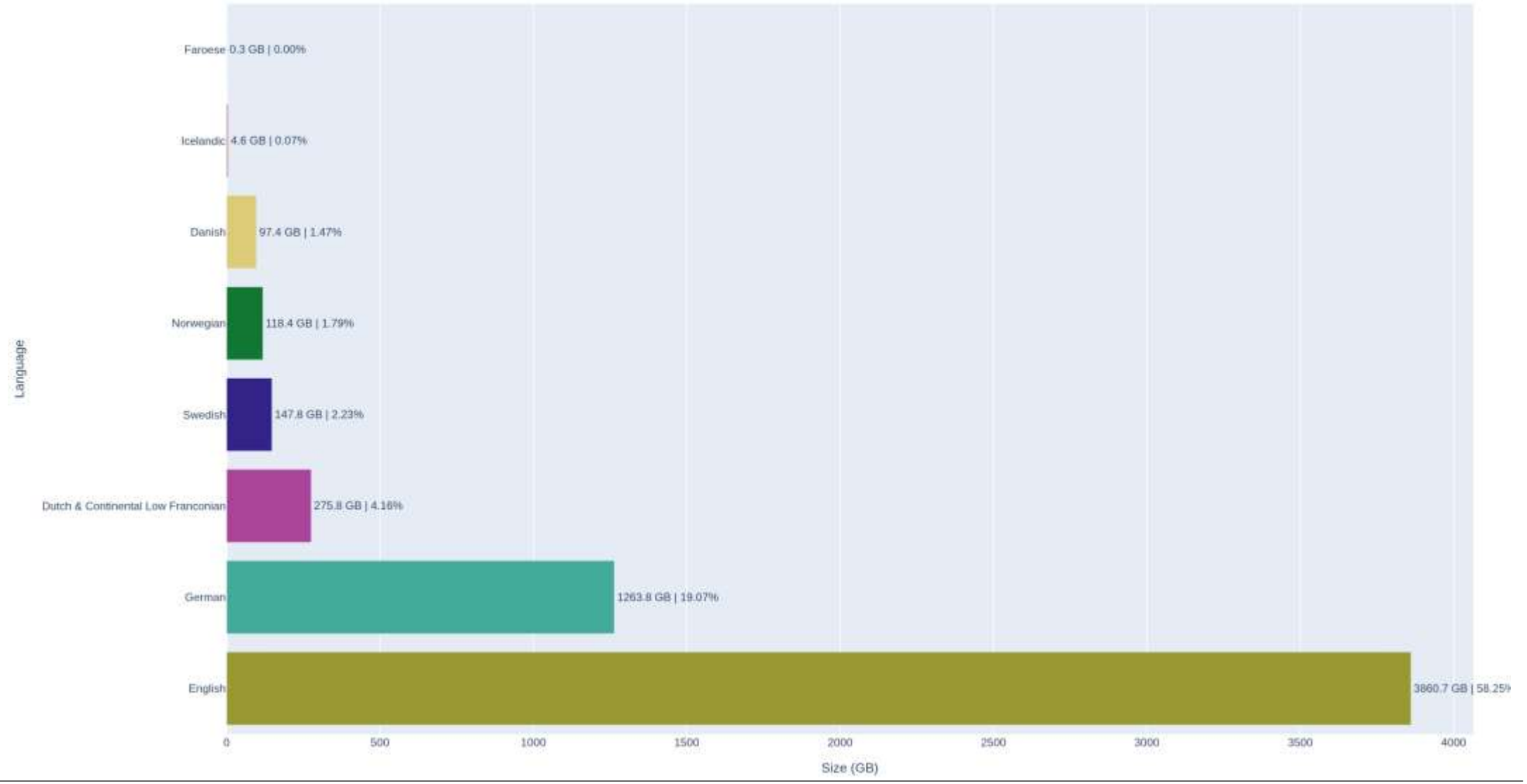
Funded by the European Union

TrustLLM Currently Curated Training Dataset (6.4TB)



Funded by
the European Union

TrustLLM Dataset Language Distribution



Funded by the European Union

Data Sources – language ratio

MultiSynt (2.2 TB Training Data)

Category	Token %
German	32.52%
English	25.92%
Norwegian	18.83%
Swedish	14.41%
Dutch	7.42%
Danish	0.83%
Icelandic	0.08%

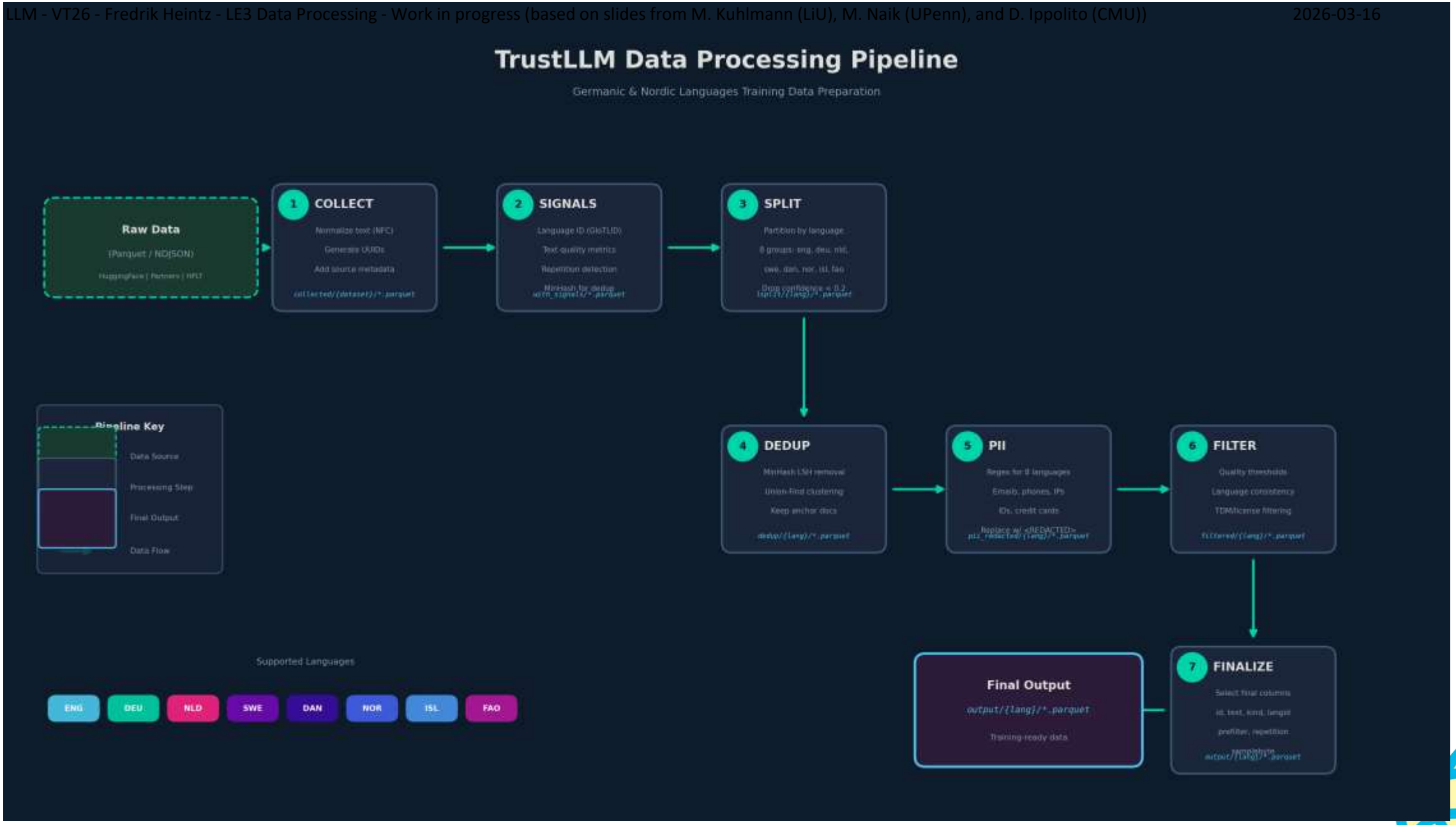
Nemotron-Pretraining-Specialized-v1 (329 GB)

Category	Token %
English	92.60%
Math	7.11%
Code	0.29%

Nemotron-CC-Math-v1 (244 GB)

Category	Token %
Math	100%

Funded by the European Union



Funded by the European Union

Pipeline - TrustLLM

Component	Tool	Purpose
Language ID	FastText (GloTLID)	Classify 30+ Germanic language variants
Text processing	Polars + polars-textproc	Streaming DataFrame operations
MinHash	polars-textproc.minhash	14-bucket locality-sensitive hashing
PII detection	pii-regexp	Country-specific regex (dan, deu, eng, fao, isl, nld, nor, swe)
Orchestration	waluigi + submitit	DAG scheduling on SLURM HPC

Funded by the European Union

PII & Article 9 handling

- Propella is an open-source ML document annotation model (ellamind) that produces structured metadata across 18 properties per document.
  - https://huggingface.co/ellamind/propella-1-4b
  - Document-level quality, safety, and compliance classification
  - PII presence detection (ML-based, complements regex)
  - Content safety scoring (safe / sensitive / nsfw / illegal)
  - Domain and audience classification for curriculum design
  - Supports 64K token context window
- Existing Resource: 3.27B row annotation dataset already published by OpenEuroLLM on HuggingFace
  - (FineWeb-2, FinePDFs, HPLT3.0, finewiki, PleIAs/SYNTH, Nemotron-CC(subset), nemotron-cc-10K-sample, German Commons)

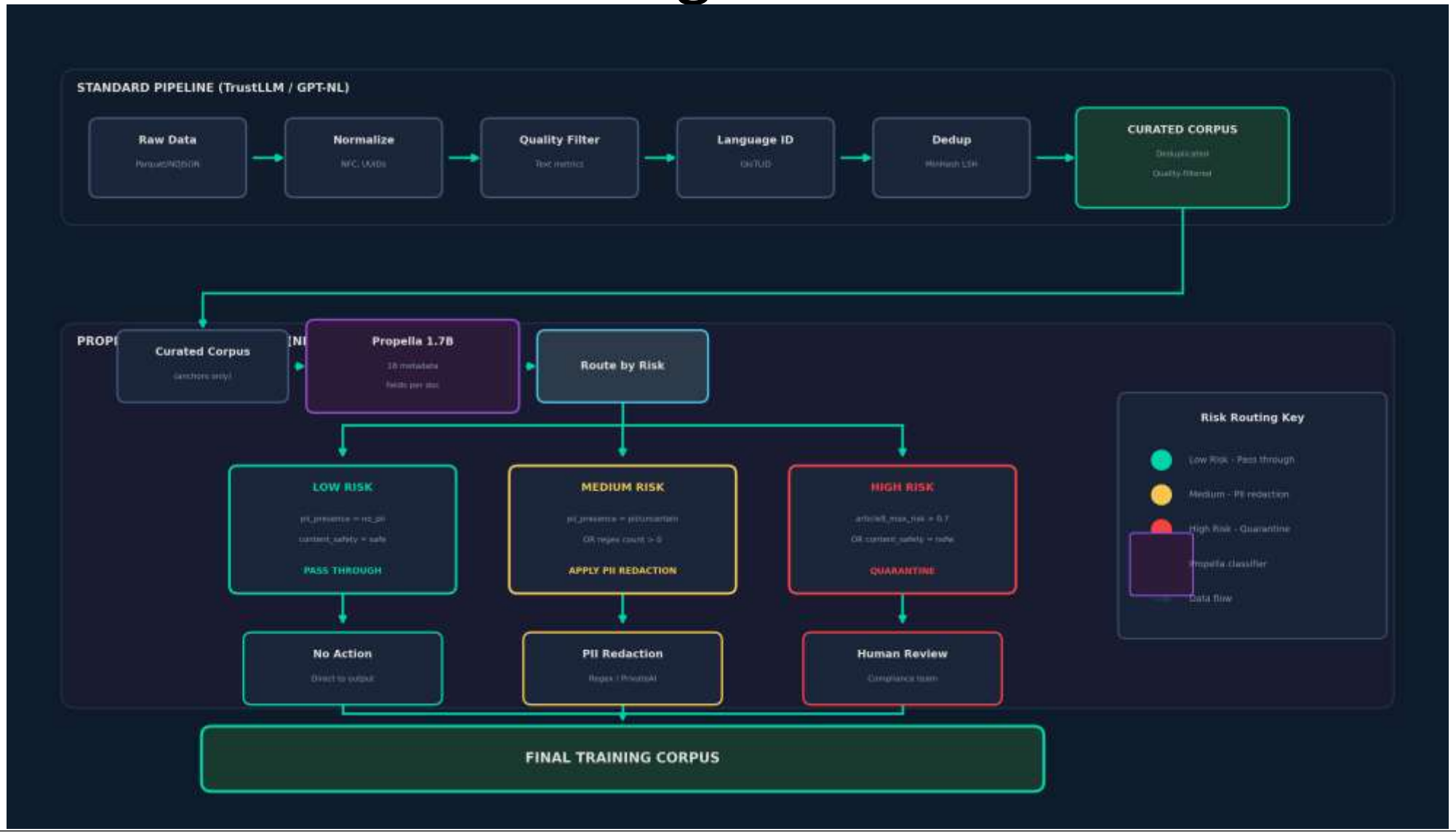
Funded by the European Union

PII & Article 9 Handling

Category	Property	Example Values
Core Content	content_integrity	intact, partial, corrupted
	text_code_ratio	text_heavy, balanced, code_heavy
	word_count_category	<100, 100-500, 500-2000, 2000-10000, >10000
Classification	description	Free-text summary (max 100 chars)
	document_type	article, documentation, forum, academic, other
	business_sector	technology, legal, medical, finance, general
	technical_depth	basic, intermediate, advanced, expert
Quality & Value	quality	excellent, good, acceptable, poor, unacceptable
	information_density	very_dense, dense, moderate, sparse
	educational_value	high, medium, low, none
	reasoning_content	substantial, moderate, minimal, none
Audience & Purpose	audience_level	general, educated, professional, expert
	commercial_bias	advertorial, sponsored, neutral, independent
	time_sensitivity	evergreen, recent, dated, obsolete
Safety & Compliance	**content_safety**	safe, sensitive, nsfw, illegal
	pii_presence	**no_pii, pii, uncertain**
Geographic	regional_relevance	global, regional, local
	country_relevance	ISO country code or "global"

Funded by the European Union

PII & Article 9 Handling



PII & Article 9 Handling

Signal	Source	Strength	Weakness
Regex PII count	pii-regexp (TrustLLM) or PrivateAI (GPT-NL)	High precision on structured PII (IBANs, phone numbers, SSNs)	Misses contextual PII (names in sentences, implied identities)
Propella pii_presence	ML model (document-level)	Catches contextual PII, understands document semantics	Document-level only (not token-level), binary signal
Propella business_sector	ML model	Flags high-risk domains (medical, legal, finance)	Indirect signal, not PII-specific
Propella content_safety	ML model	Identifies sensitive/nsfw/illegal content	Broad categories, not PII-specific

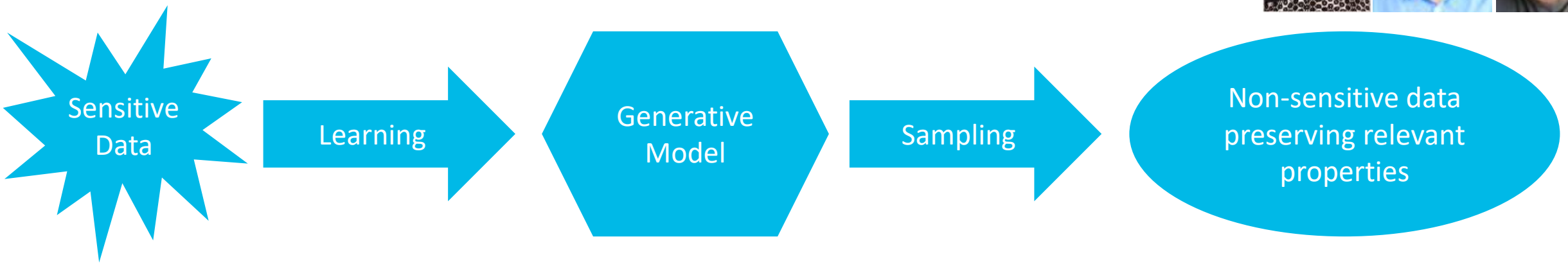
Regex Count	Propella PII Flag	Article 9 Risk	Decision
0	no_pii	< 0.7	**CLEAN** -- pass through
0	pii or uncertain	< 0.7	**REDACT** -- apply PII removal
> 0	no_pii	< 0.7	**FLAG** -- likely false positive, redact conservatively
> 0	pii	< 0.7	**REDACT** -- confirmed PII, apply full redaction
any	any	>= 0.7	**REVIEW** -- route to human compliance review
any	any	illegal/nsfw	**QUARANTINE** -- exclude from training

Funded by the European Union

Privacy-preserving synthetic data generation

[R. Ramachandranpillai, Md F. Sikder, D. Bergström]

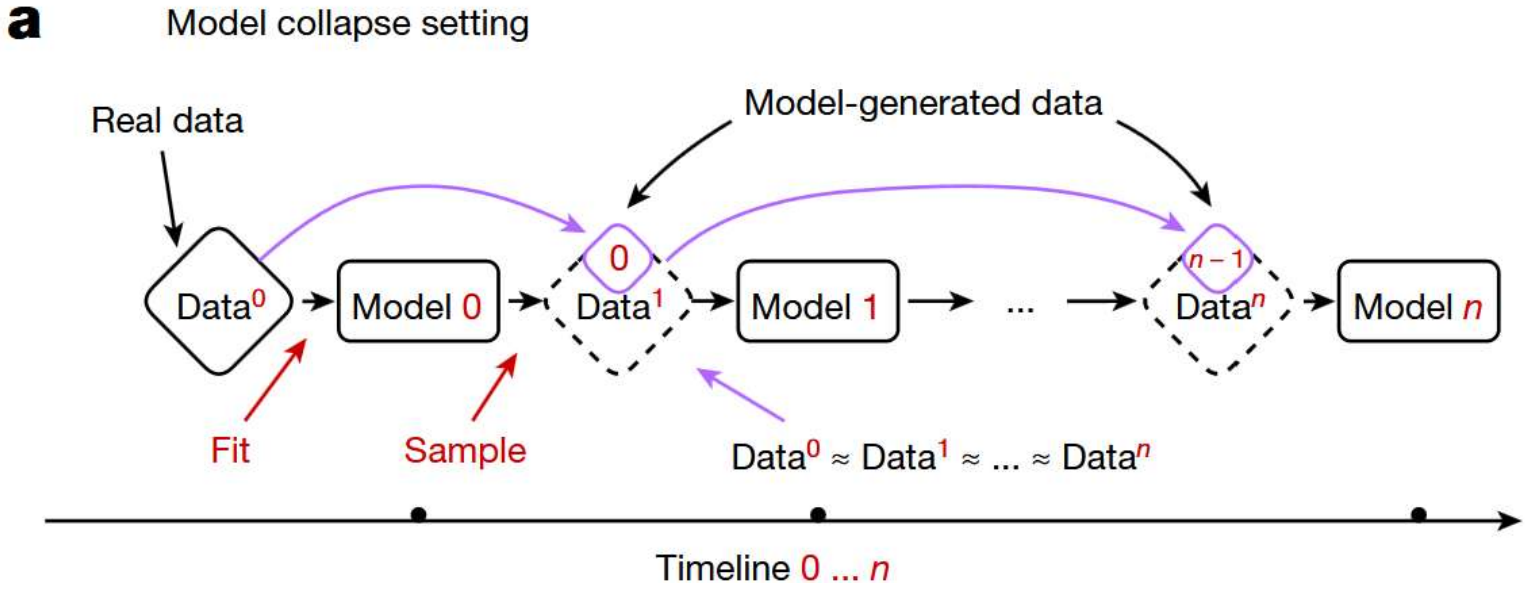


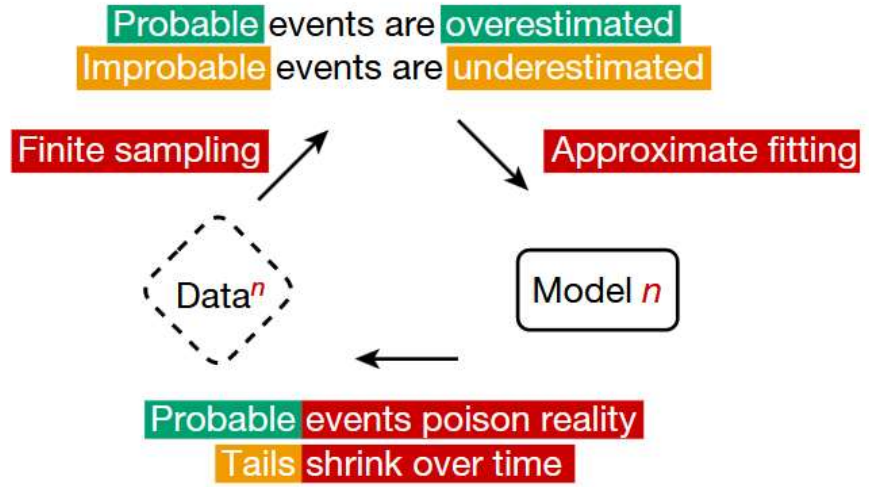


1. Learn a generative model that captures the probability distribution of the sensitive data
2. Create a synthetic data set from the generative model that both captures the salient features of the original data set **and** is non-sensitive
3. Methods for verifying that the synthetic data set is accurate enough
4. Methods for verifying that the synthetic data set is non-sensitive

Fair Latent Deep Generative Models (FLDGMs) for Syntax-Agnostic and Fair Synthetic Data Generation, *Resmi Ramachandranpillai*, Md Fahim Sikder*, Fredrik Heintz*, ECAI23
Bt-GAN: Generating Fair Synthetic Healthdata via Bias-transforming Generative Adversarial Networks, *Resmi Ramachandranpillai, Md Fahim Sikder, David Bergström, Fredrik Heintz,* Accepted to JAIR.

Model Collapse Using Only Synthetic Data

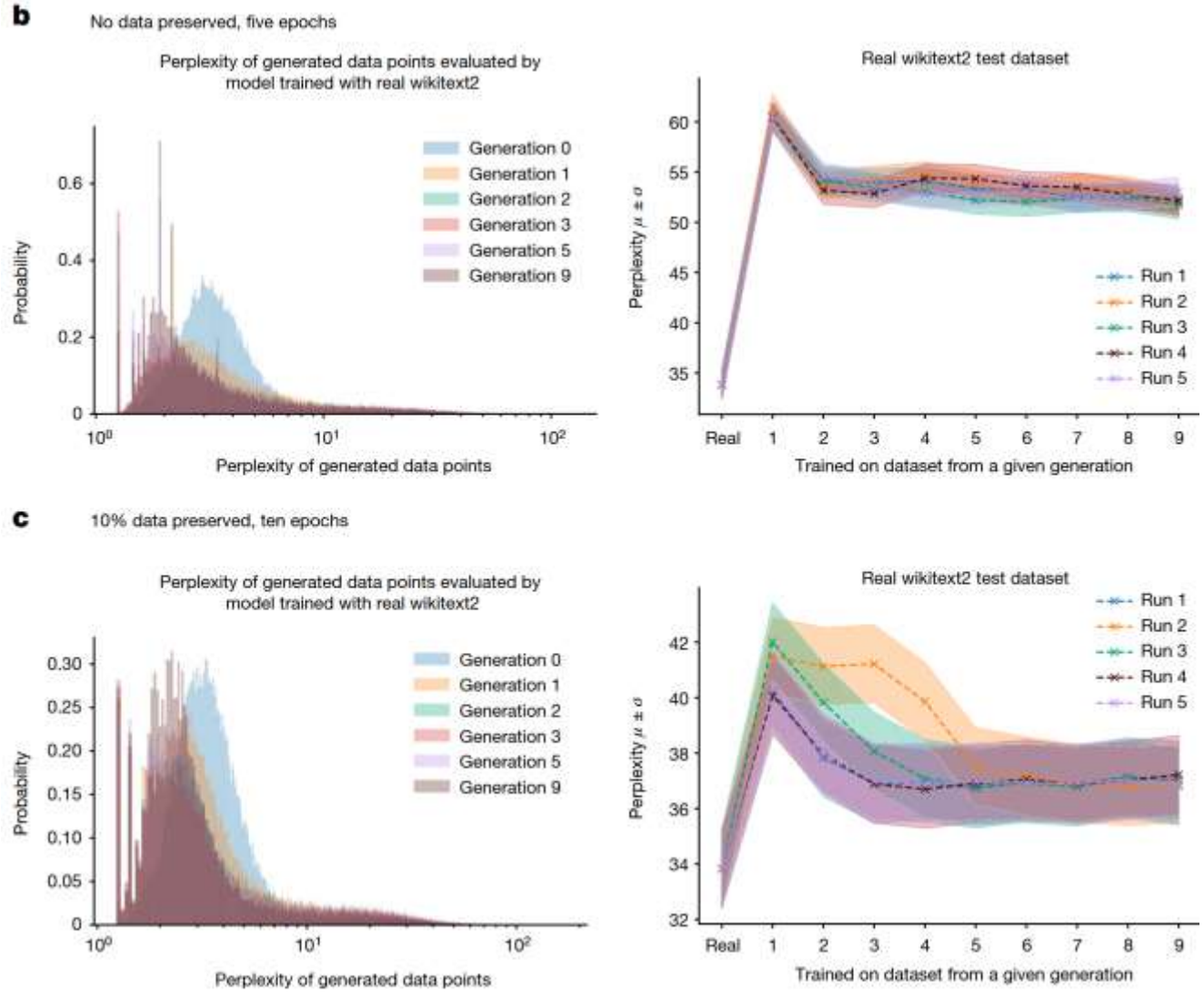




Shumailov, I., Shumaylov, Z., Zhao, Y. *et al.* AI models collapse when trained on recursively generated data. *Nature* **631**, 755–759 (2024). https://doi.org/10.1038/s41586-024-07566-y

Funded by the European Union

Model Collapse Using Only Synthetic Data



Shumailov, I., Shumaylov, Z., Zhao, Y. *et al.* AI models collapse when trained on recursively generated data. *Nature* **631**, 755–759 (2024). https://doi.org/10.1038/s41586-024-07566-y

Funded by the European Union

Data Mixture

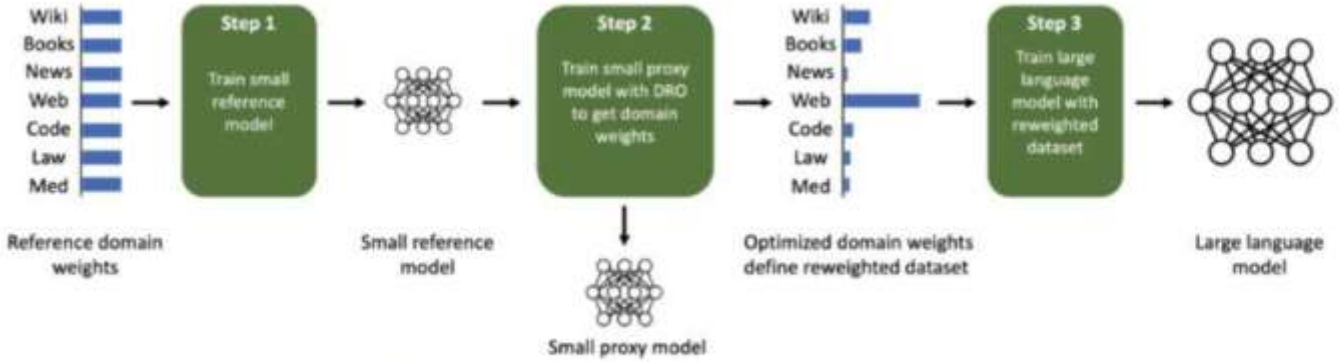
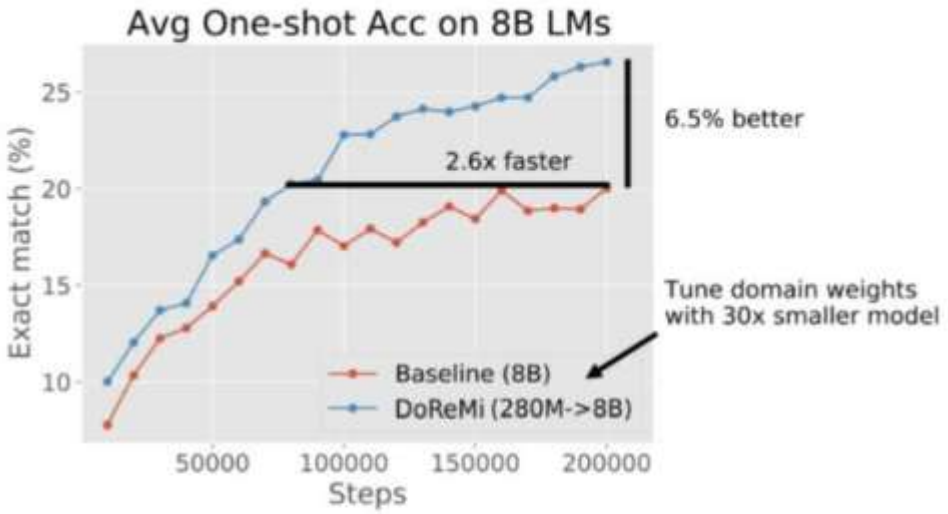


Figure 1: Given a dataset with a set of domains, Domain Reweighting with Minimax Optimization (DoReMi) optimizes the domain weights to improve language models trained on the dataset. First, DoReMi uses some initial reference domain weights to train a reference model (Step 1). The reference model is used to guide the training of a small proxy model using group distributionally robust optimization (Group DRO) over domains (Nemirovski et al., 2009, Oren et al., 2019, Sagawa et al., 2020), which we adapt to output domain weights instead of a robust model (Step 2). We then use the tuned domain weights to train a large model (Step 3).



Xie, S.M., Pham, H., Dong, X., Du, N., Liu, H., Lu, Y., Liang, P., Le, Q.V., Ma, T. and Yu, A.W., 2023. DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining. arXiv preprint arXiv:2305.10429.

Funded by the European Union

Data Order

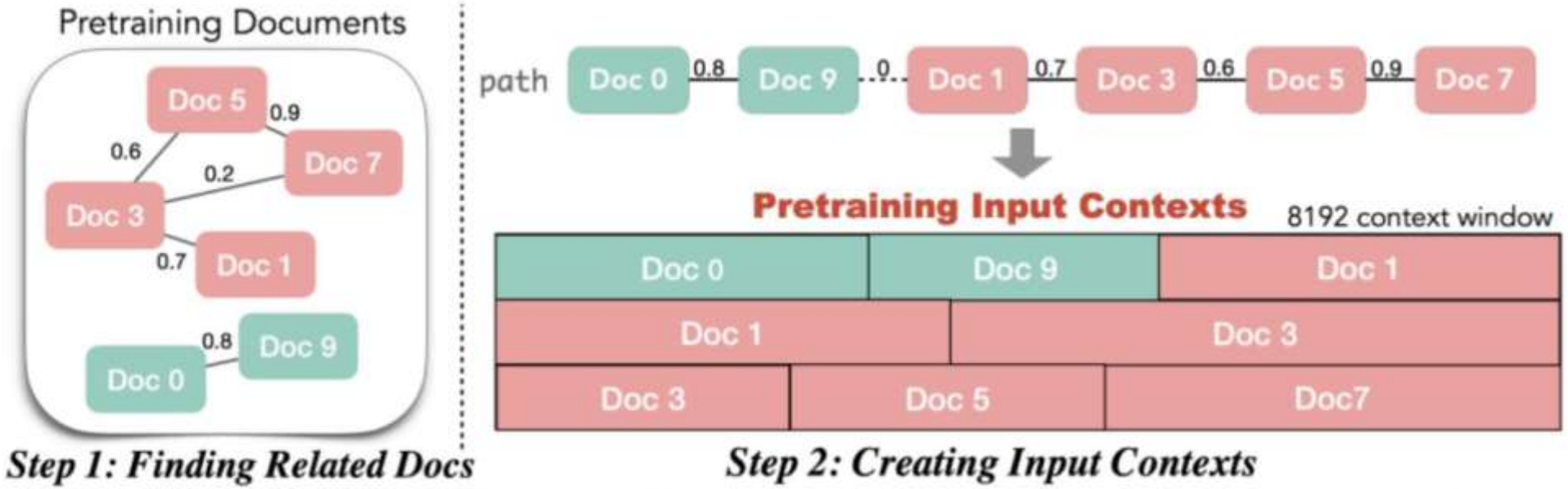


Figure 2: **Illustration of In-Context Pretraining.** In-Context Pretraining first finds related documents at scale to create a document graph (§2.1) and then builds pretraining input contexts by traversing the document graph (§2.2). Along the path, documents are concatenated into a sequence and subsequently divided to form fixed-sized input contexts (e.g., 8192 token length).

Shi, W., Min, S., Lomeli, M., Zhou, C., Li, M., Lin, V., Smith, N.A., Zettlemoyer, L., Yih, S. and Lewis, M., 2023. In-Context Pretraining: Language Modeling Beyond Document Boundaries. arXiv preprint arXiv:2310.10638.

Funded by the European Union

Data Masking

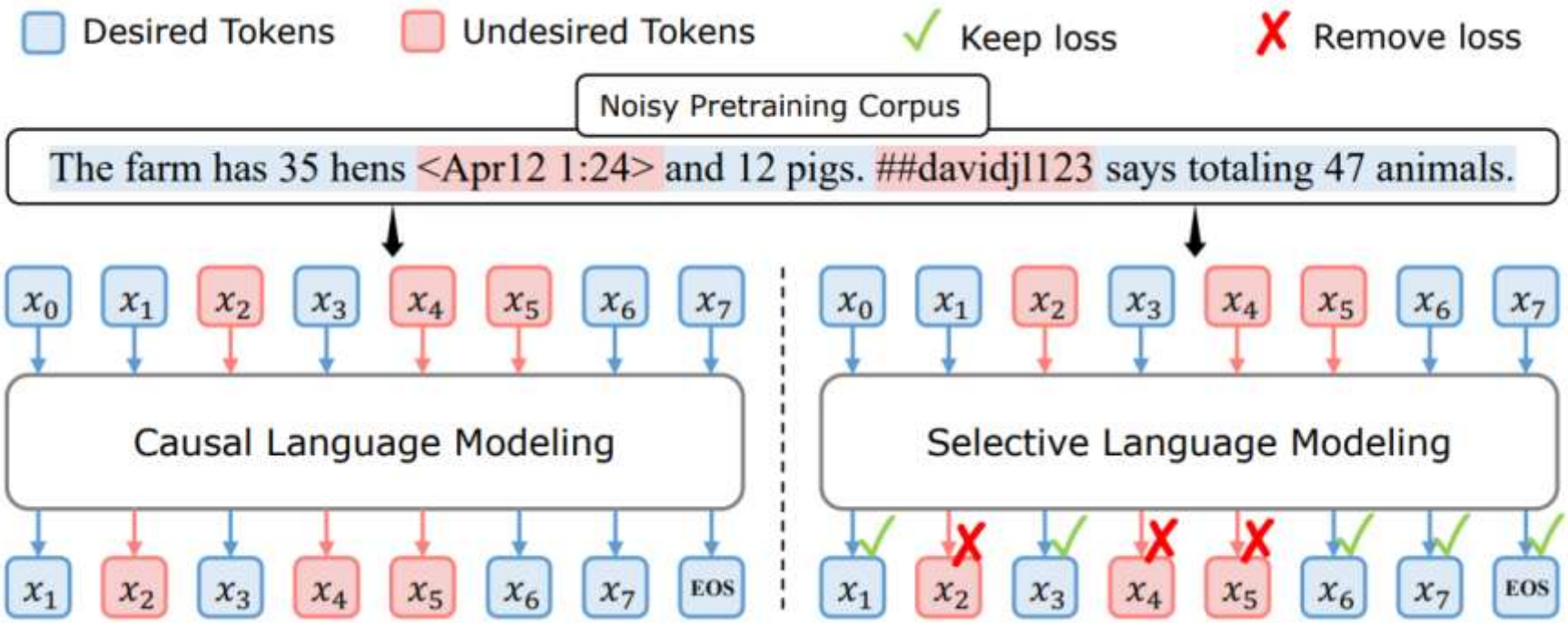


Figure 2: **Upper:** Even an extensively filtered pretraining corpus contains token-level noise. **Left:** Previous Causal Language Modeling (CLM) trains on all tokens. **Right:** Our proposed Selective Language Modeling (SLM) selectively applies loss on those useful and clean tokens.

RHO-1: Not All Tokens Are What You Need. https://arxiv.org/pdf/2404.07965

Funded by the European Union

Data Scale Matters

https://epoch.ai/data-insights/dataset-size-trend

Funded by the European Union

Pre-Training Data Quality Reduces Compute Needs

Recent work finds smaller amounts of higher quality data removes the need for a larger model.

There is increasing evidence that efforts to better curate training corpus, including **deduping, pruning data and investing in synthetic data** can compensate for the need for larger networks and/or improve training dynamics.

	% train examples with dup in train	dup in valid	% valid with dup in train
C4	3.04%	1.59%	4.60%
RealNews	13.63%	1.25%	14.35%
LM1B	4.86%	0.07%	4.92%
Wiki40B	0.39%	0.26%	0.72%

Table 2: The fraction of examples identified by NEARDUP as near-duplicates.

Lee et al. 2022

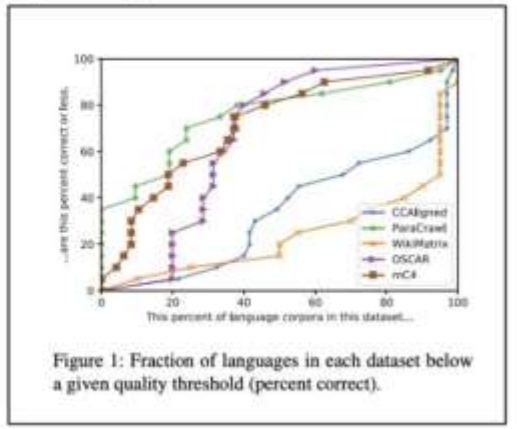
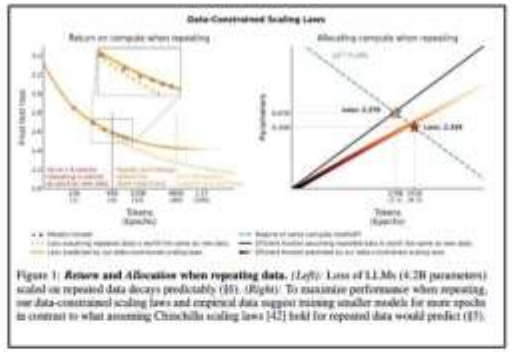
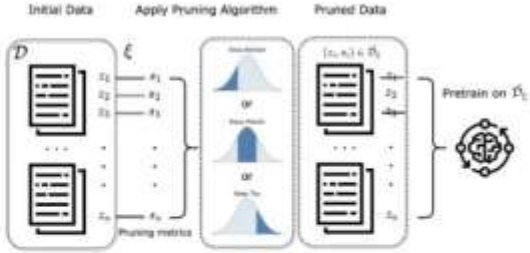


Figure 1: Fraction of languages in each dataset below a given quality threshold (percent correct).

Kreutzer at al. 2022



Muennighoff et al. 2023



Marion et al. 2023

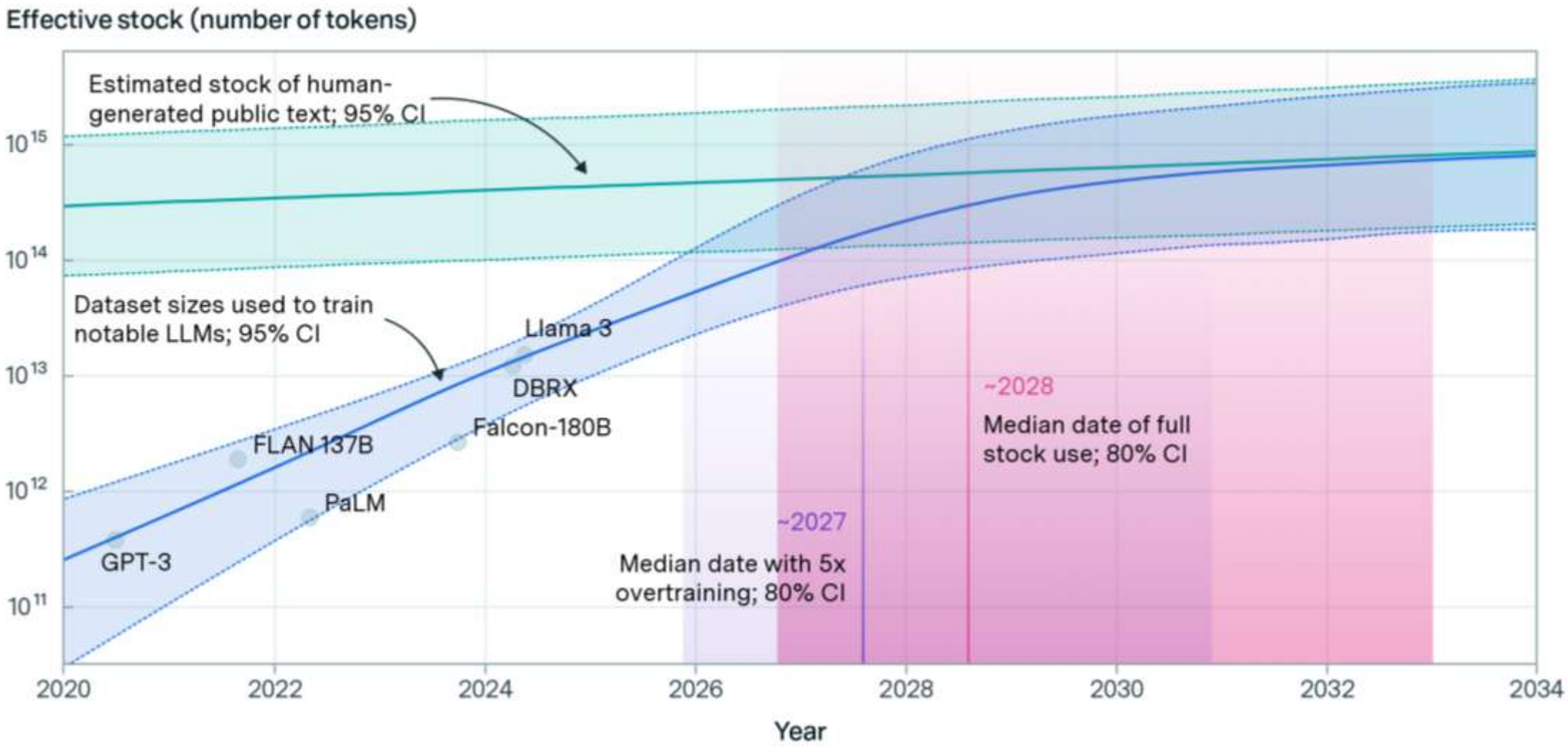
Cohere For AI

S. Hooker. On the Limitations of Compute Thresholds as a Governance Strategy. 2024.

Funded by the European Union

Projections of the stock of public text and data usage

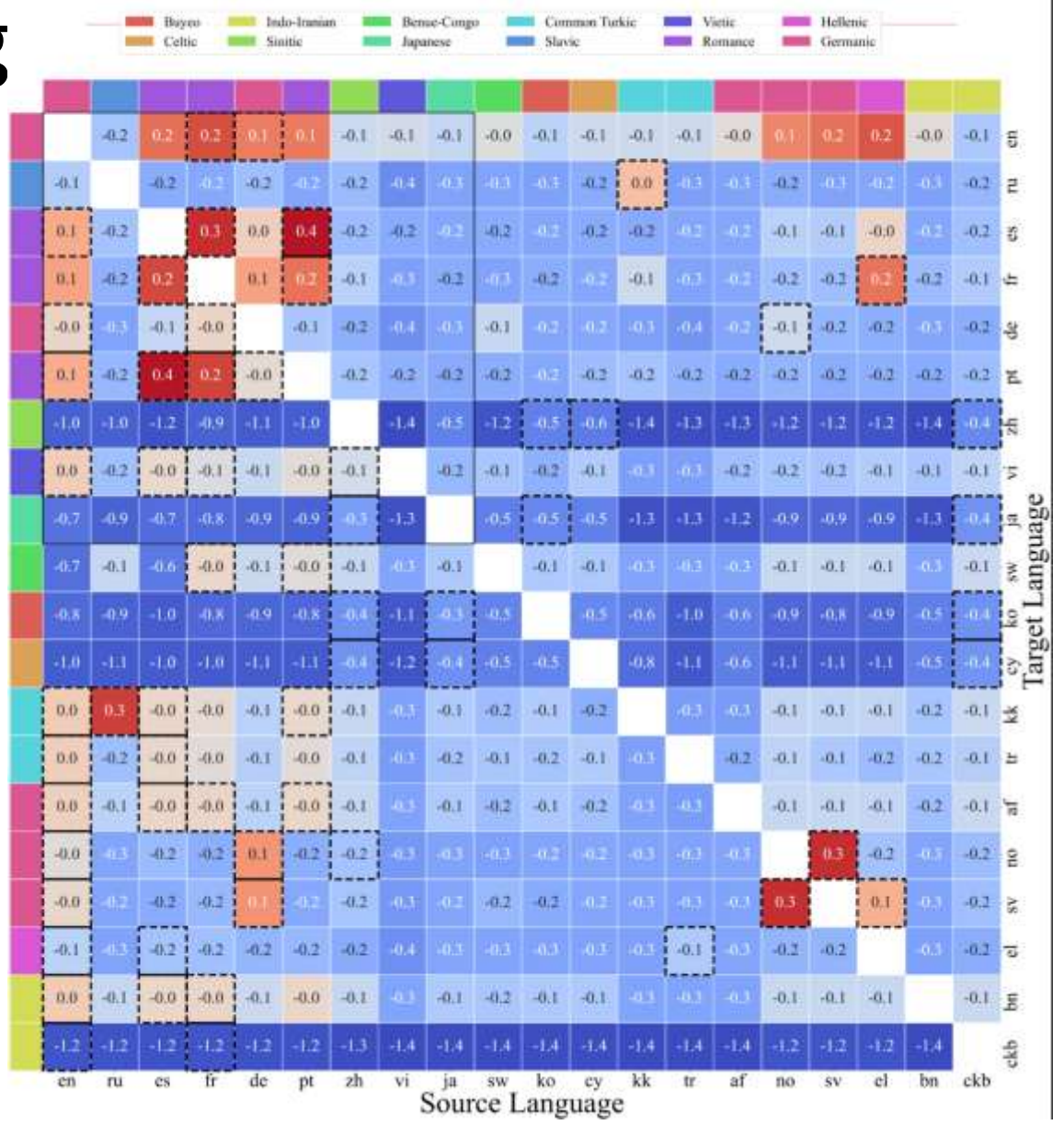
EPOCH AI



https://epoch.ai/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data

Funded by the European Union

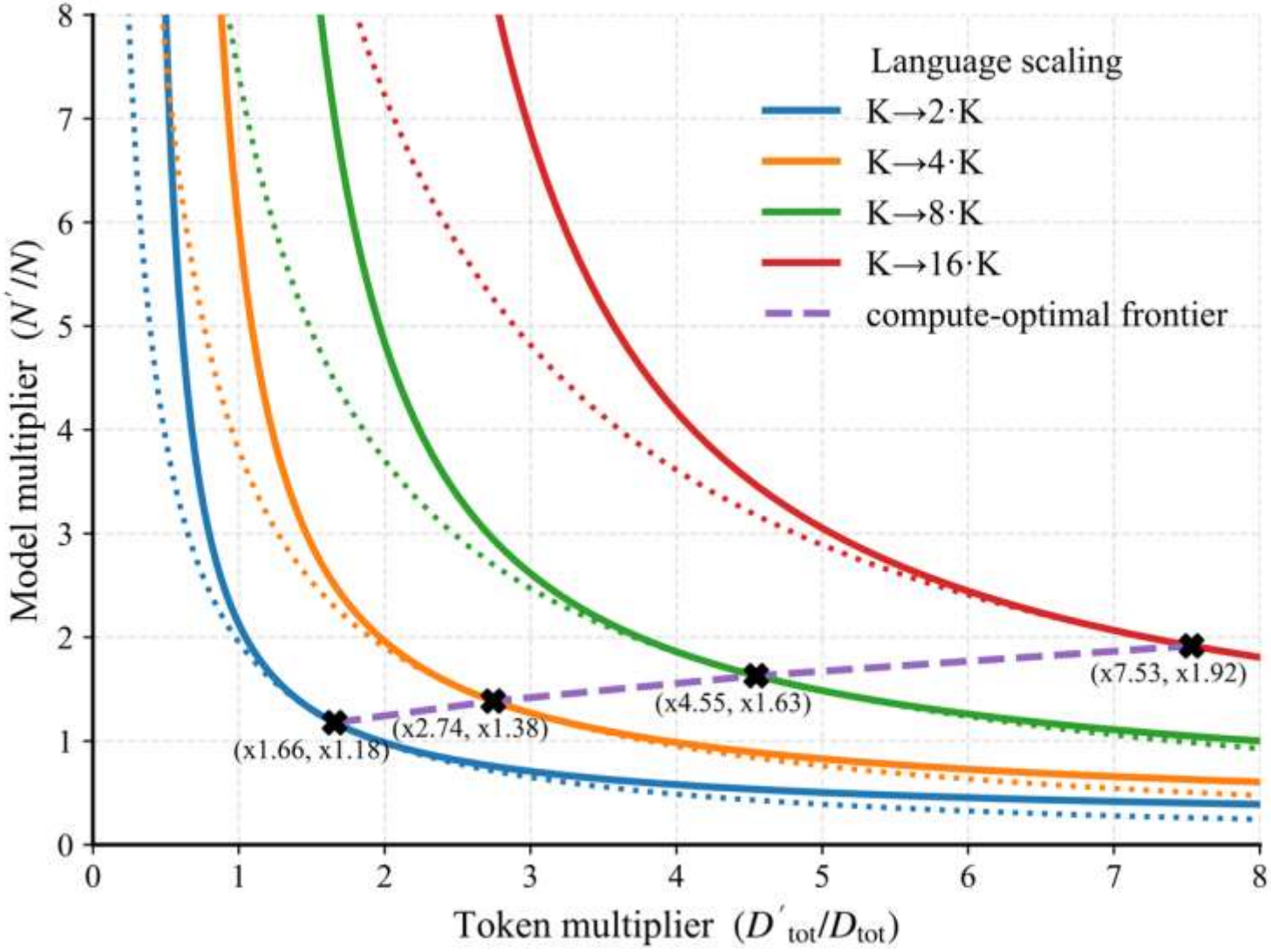
ATLAS: Practical scaling laws for multilingual models



https://research.google/blog/atlas-practical-scaling-laws-for-multilingual-models/

Funded by the European Union

ATLAS: The “curse of multilinguality”



https://research.google/blog/atlas-practical-scaling-laws-for-multilingual-models/

Funded by the European Union

- LAIM LE3 VT2026:

Tokenization
Data Processing Pipeline

www.ida.liu.se/~freho08/llm

LINKÖPING UNIVERSITY