

✨ ✨ **Tik-to-tok** ✨ ✨

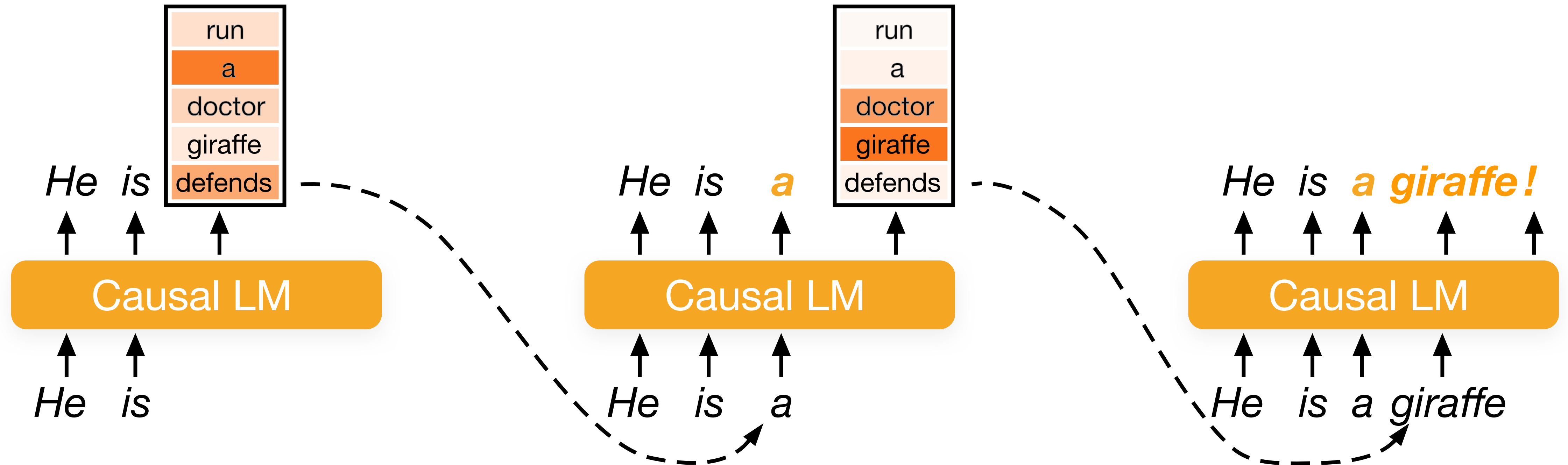
Turn your monolingual LLMs into
multilingual maven

Pieter Delobelle & François Remy

February 27, 2024

Chapter 1: Tik... to Tok?

LLMs: generating text token per token



Tokens and embeddings

No, I am not a giraffe.

Tokens and embeddings

No, I am not a giraffe.



No, I am not a giraffe.

Tokens and embeddings

No, I am not a giraffe.



No, I am not a giraffe.



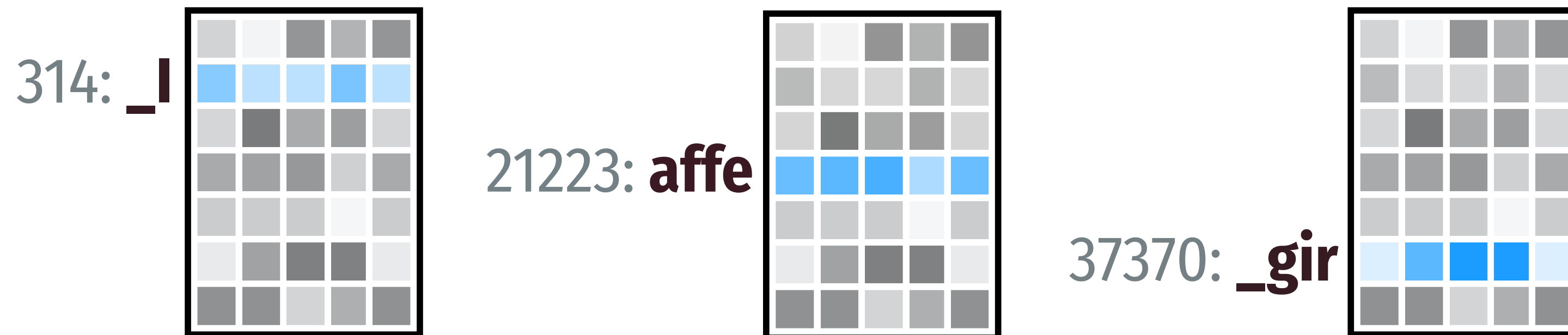
[2822, 11, 358, 1097, 539, 264, 41389, 38880, 13]

Tokens and embeddings

No, I am not a giraffe.

No, I am not a giraffe.

[2822, 11, 358, 1097, 539, 264, 41389, 38880, 13]



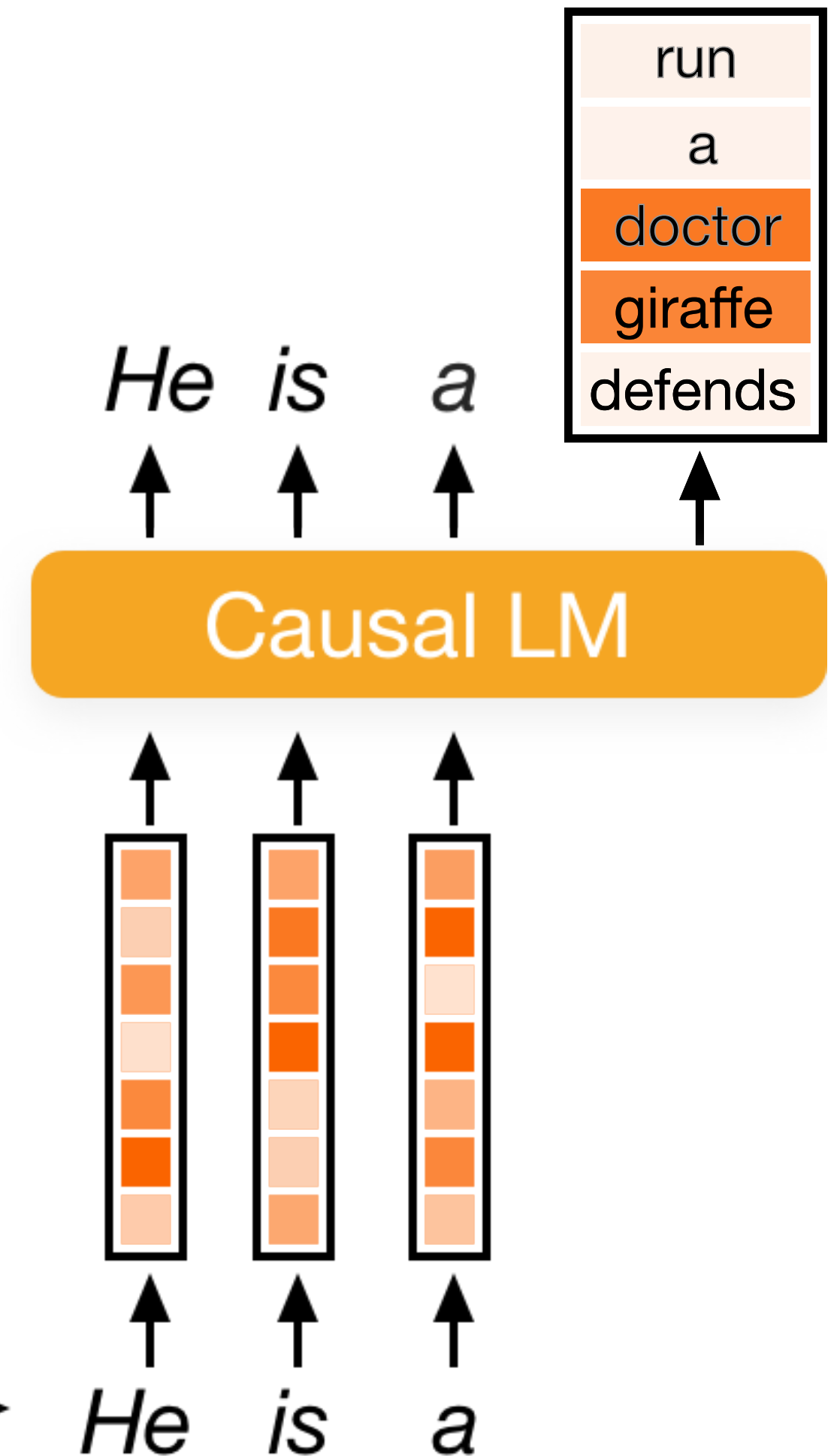
How does a LM learn that?

It is the tallest living terrestrial animal.

Giraffes live in herds.

He is a giraffe.

IUCN recognises one species of giraffe.



Pretraining is expensive, but worth it



One book
40-50k tokens

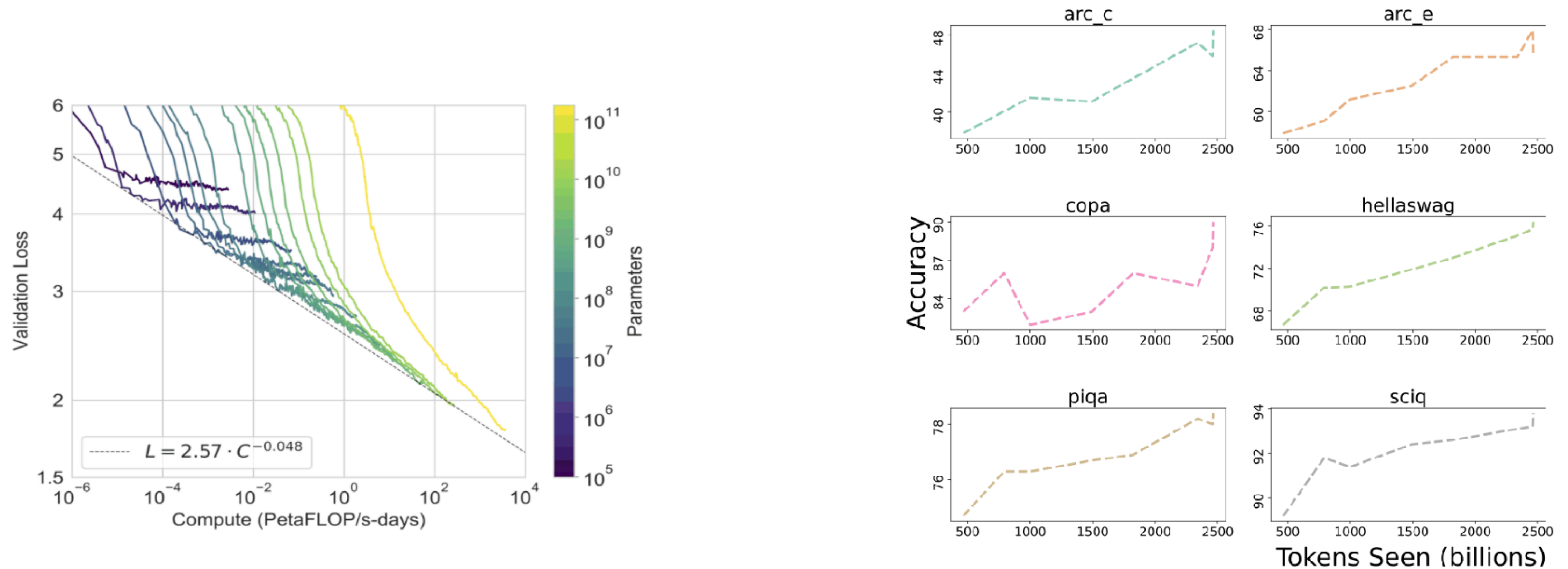


One bookshelf
1.6M - 2.5M tokens



One LLM training set
2.5T - 6T tokens
~2 500 000 bookshelves

Pretraining is expensive, but worth it



Chance of generating Franken-Dutch



Cost

Dutch tokenizer

Training from scratch
GPT-NL... one day?

?

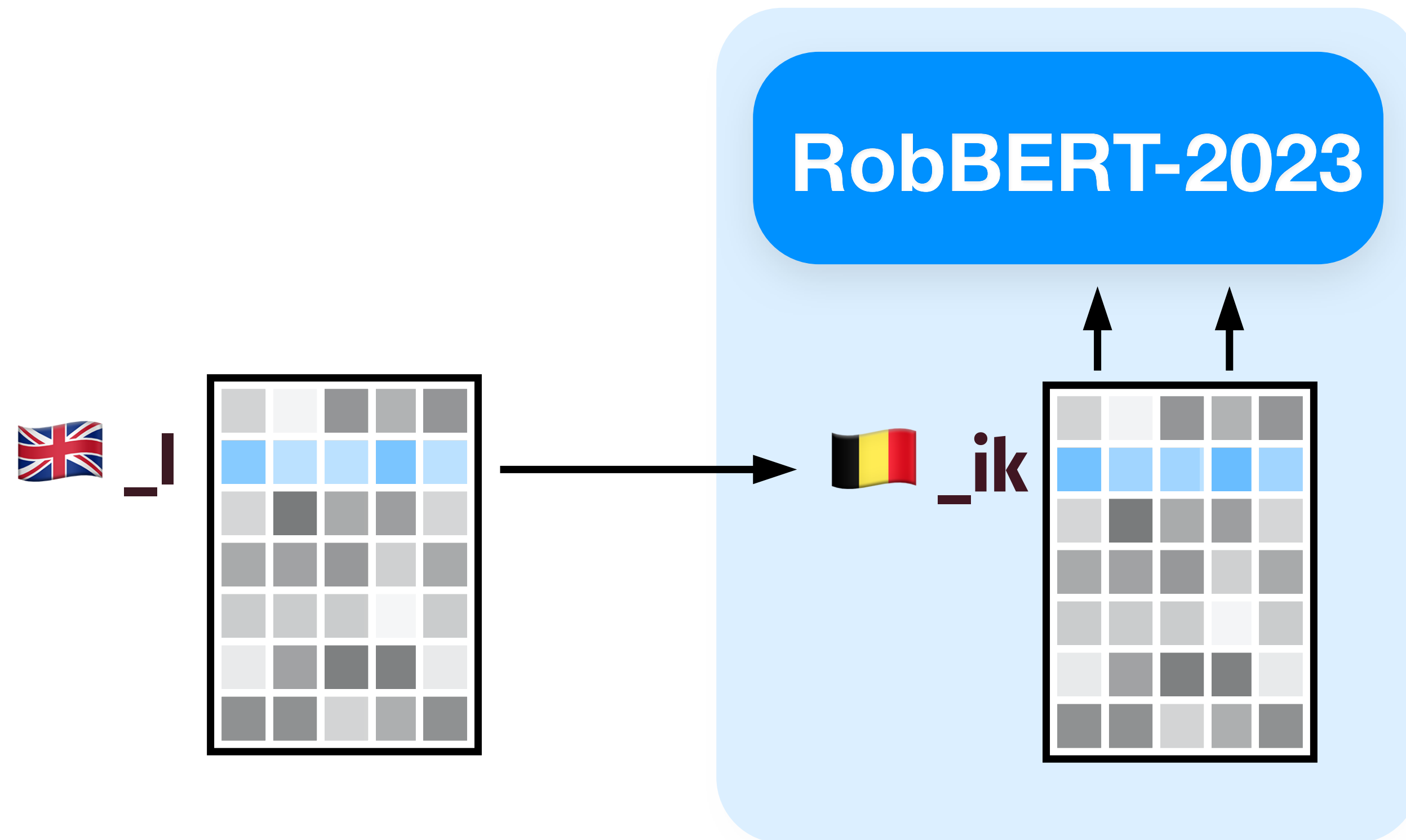
English tokenizer

Finetuning
GEITje, ...

LoRA finetuning
"BLOOM-NL", ...

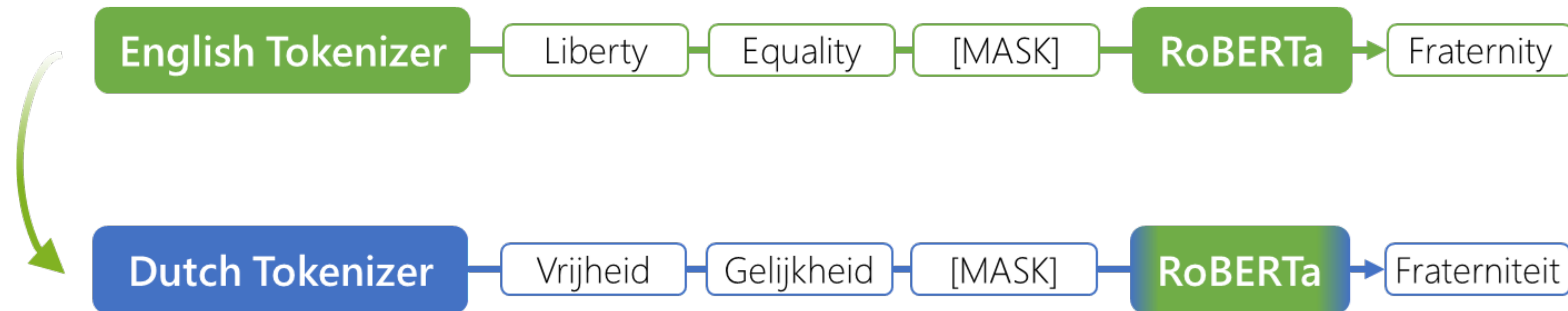
Prompting
Mistral with Dutch prompts

How can we reuse these models?



Tik-to-tok

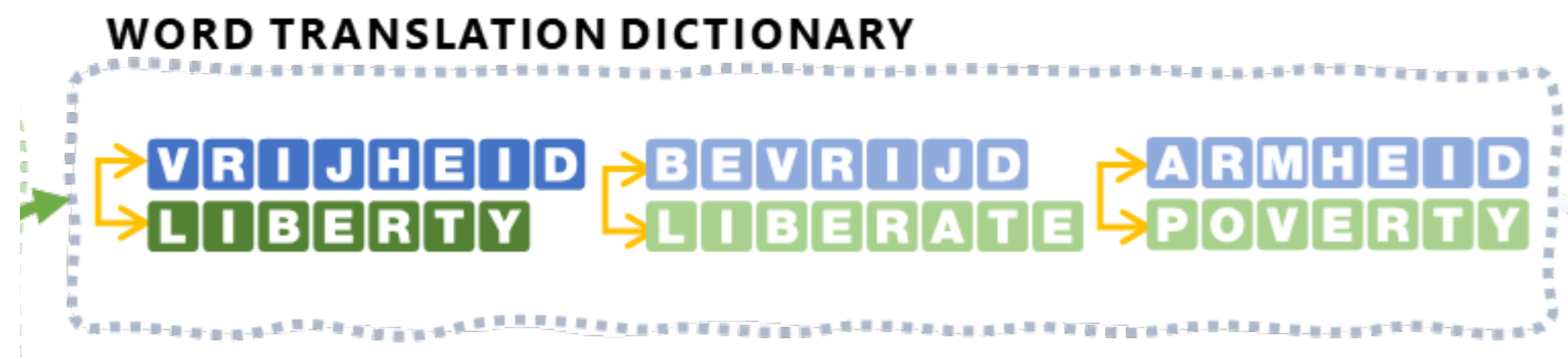
Our objective:



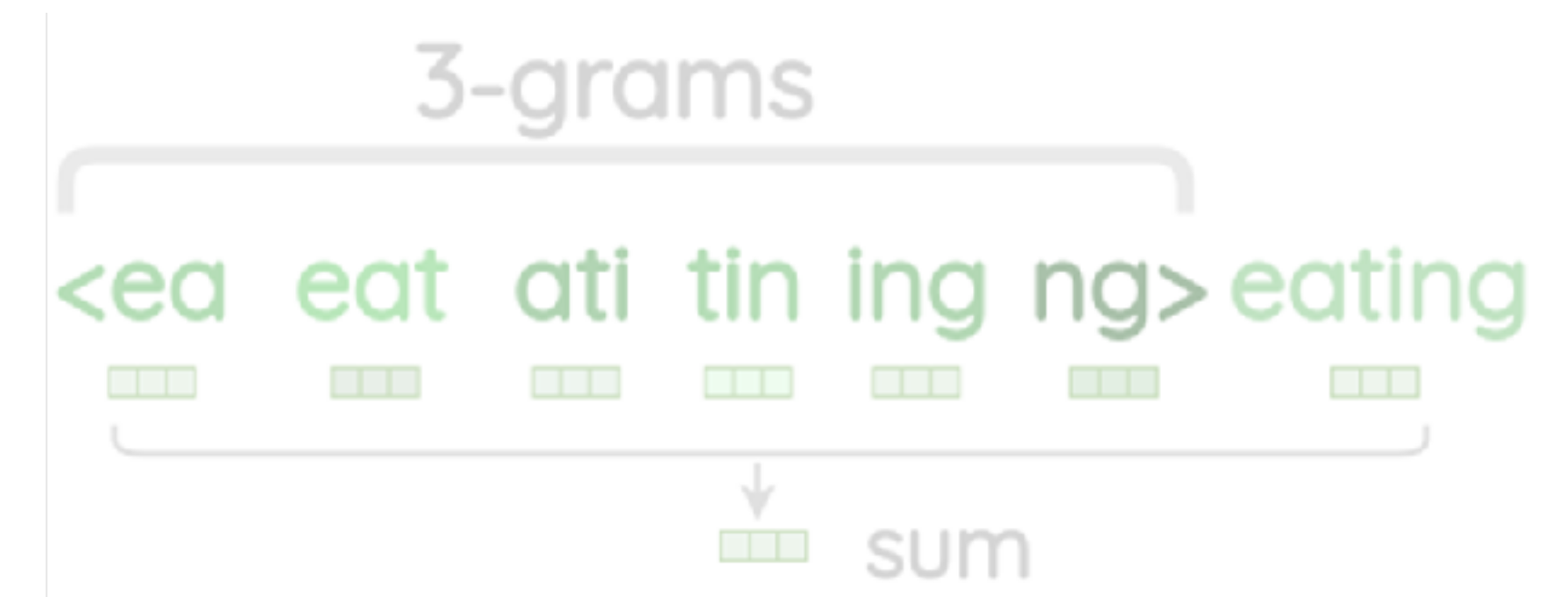
Tik-to-tok

Our tools:

1. Good old translation dictionaries!



2. Character n-gram embeddings



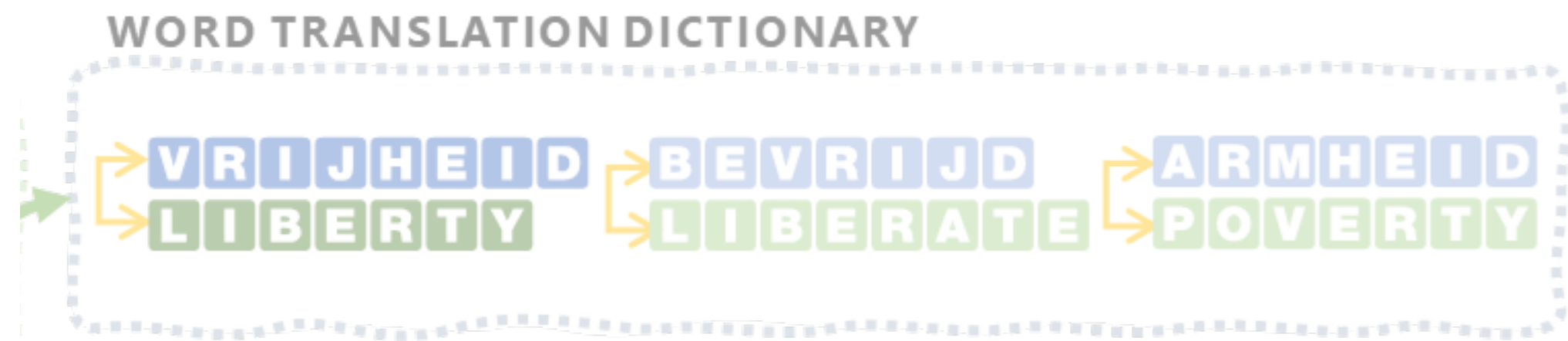
Using FastText, you can compute embeddings for incomplete words, by summing the embedding of the character n-grams it contains.

This enables to embed tokens!

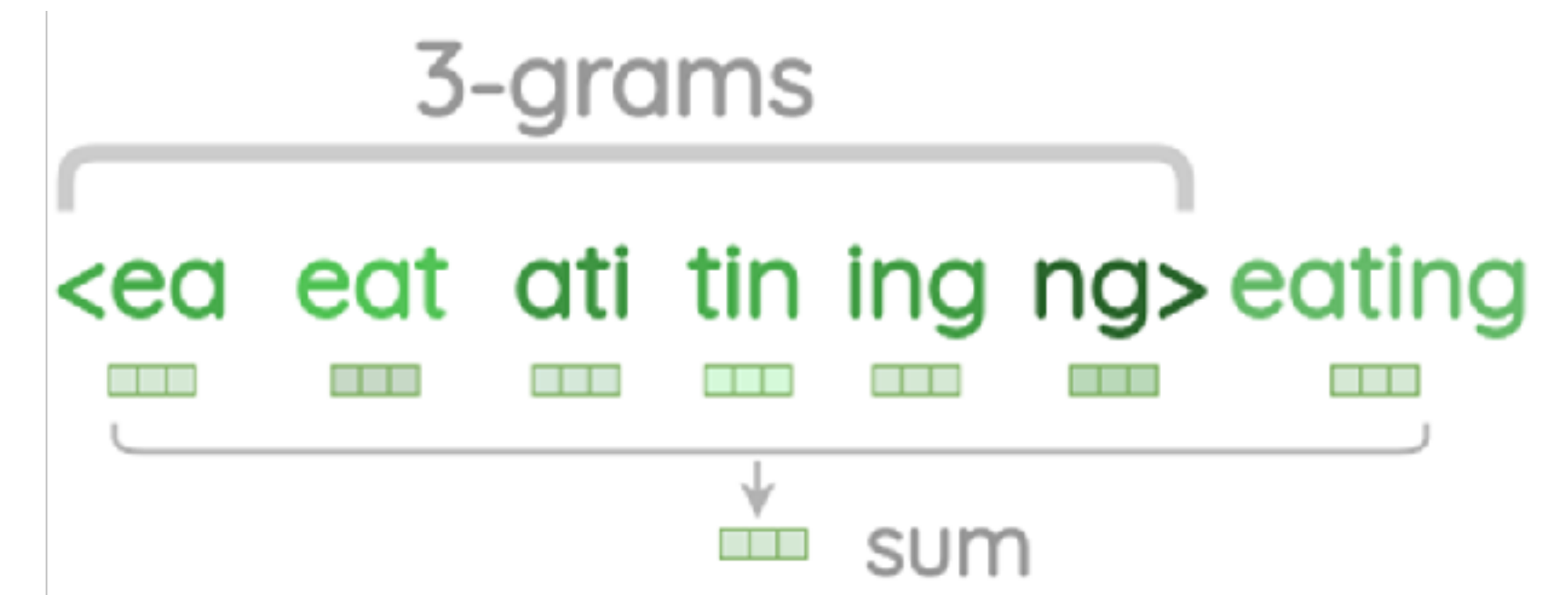
Tik-to-tok

Our tools:

1. Good old translation dictionaries!



2. Character n-gram embeddings



Using FastText, you can compute embeddings for incomplete words, by summing the embedding of the character n-grams it contains.

This enables to embed tokens!

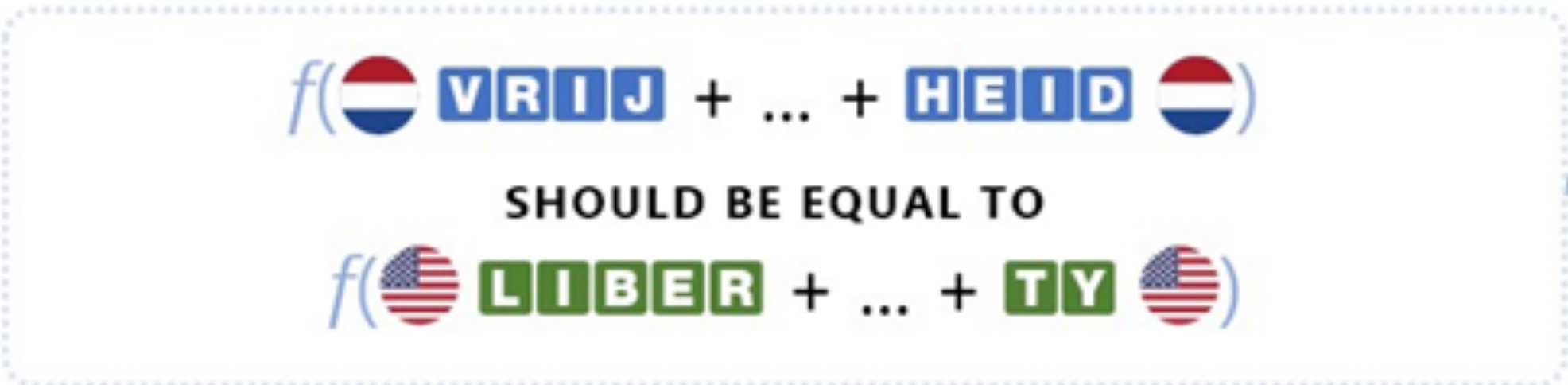
Tik-to-tok

Our tools:

1. Good old translation dictionaries!



(3) FAST-TEXT WITH SKIPGRAM OBJECTIVE



(4) TOKEN EMBEDDING + SOURCE NEIGHBORS RETRIEVAL



2. Character n-gram embeddings













Using FastText, you can compute embeddings for incomplete words, by summing the embedding of the character n-grams it contains.

This enables to embed tokens!

Chapter 2: BERT

RobBERT-2023



Model	Avg RER	Word		Sentence Pair		Document		
		POS RER _{Acc}	NER RER _{F1}	CR RER _{Acc}	NLI RER _{Acc}	SA RER _{Acc}	ALD RER _{F1}	QA RER _{F1}
 RobBERT ₂₀₂₃ large	18.6	15.9 _{98.1}	19.0 _{88.8}	47.1 _{79.9}	27.7 _{89.3}	21.9 _{94.7}	16.5 _{65.6}	16.2 _{75.1}
 DeBERTa _{v3} large	15.7	17.9 _{98.2}	10.9 _{87.6}	35.4 _{75.4}	24.1 _{88.7}	-6.4 _{92.8}	12.5 _{64.0}	48.4 _{84.7}
 XLM-R _{large}	14.4	26.5 _{98.4}	29.7 _{90.3}	-25.8 _{52.2}	24.4 _{88.8}	13.2 _{94.2}	19.0 _{66.6}	37.2 _{81.4}
 mDeBERTa _{v3} base	12.8	18.2 _{98.2}	17.2 _{88.5}	19.7 _{69.5}	25.2 _{88.9}	3.3 _{93.5}	12.4 _{63.9}	29.2 _{79.0}
 Tik-to-Tok _{large}	12.3	22.2 _{98.3}	24.2 _{89.5}	2.4 _{62.9}	29.5 _{89.6}	30.5 _{95.3}	-8.0 _{55.5}	41.7 _{82.7}
 Tik-to-Tok _{base}	5.7	15.1 _{98.1}	4.7 _{86.8}	6.5 _{64.5}	9.7 _{86.6}	4.1 _{93.6}	12.0 _{63.7}	17.8 _{75.6}
 RobBERT ₂₀₂₃ base	4.5	8.8 _{98.0}	6.5 _{87.0}	2.6 _{63.0}	7.8 _{86.3}	0.8 _{93.3}	7.0 _{61.7}	-5.9 _{68.6}
 RobBERT ₂₀₂₂ base	3.6	17.3 _{98.2}	7.6 _{87.2}	-10.1 _{58.2}	3.1 _{85.6}	4.0 _{93.5}	18.9 _{66.6}	-0.2 _{70.3}
 RobBERT _{v2} base	1.6	16.2 _{98.2}	4.1 _{86.7}	-10.2 _{58.1}	-3.8 _{84.6}	-0.5 _{93.2}	12.0 _{63.7}	2.2 _{71.0}
 BERTje _{base}	0.0	0.0 _{97.8}	0.0 _{86.1}	0.0 _{62.0}	0.0 _{85.2}	0.0 _{93.3}	0.0 _{58.8}	0.0 _{70.3}

Chapter 3: GPT

Low-Resource & GPT?

- **Academic feedback on RobBERT-2023:**
 - Dutch is not low-resourced enough; too much data
 - BERT models are overshadowed by GPT models



Specificities of GPT

- **Generation is more difficult than comprehension**
- **So, models are very very large!**
 - *universities can't usually afford them*
- **And they make trade-offs, like...**
 - *requiring enormous amount of text to train*
 - *dividing words into more tokens*

Larger.. language models

- **Advantages:**

- Our technique is even more valuable
- The resulting models can achieve way more

- **Disdvantages:**

- It costs much more to retrain, and takes longer



Smaller.. tokens

- **Advantages:**

- This mitigates a bit the cost of training...
- but only if we also adopt a smaller vocabulary

- **Disdvantages:**

- Mapping becomes even more difficult, because tokens / groups of letters become shared by many more words

Changes to Tik-to-Tok

- **Before:**

- We would map tokens based on the character sequences they contain.

- **After:**

- We map the tokens based on which words they are actually used in.

- For this, we use SMT.

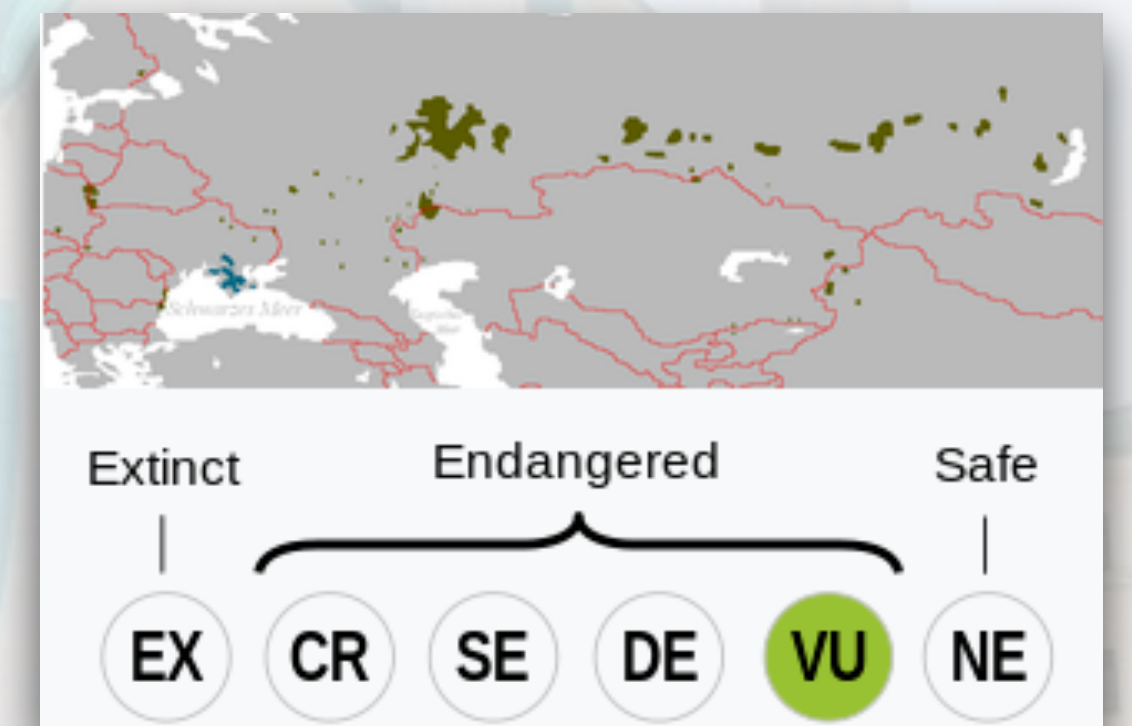
Low-Resource & GPT?

- **Academic feedback on RobBERT-2023:**
 - Dutch is not low-resourced enough; too much data
 - BERT models are overshadowed by GPT models



Tatar GPT

- **Tatar?**
 - Low-Resourced Language
 - Not an Indo-European Language
 - Not written using the Latin Alphabet*

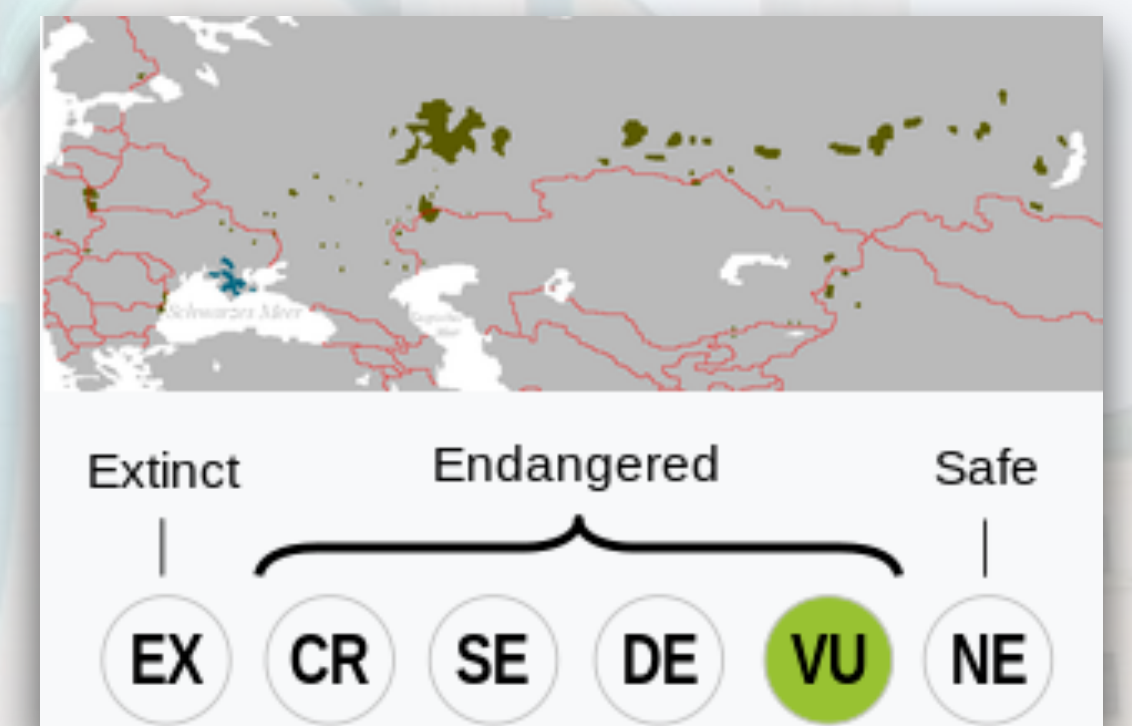


Tatar GPT

- **Tatar?**
 - Low-Resourced Language
 - Not an Indo-European Language
 - Not written using the Latin Alphabet*



Alfiya Khabibullina



What can achieve with this?

- **What we have:**

- Small but not minuscule corpus of monolingual text
- Small and mostly Bible- and Quran-based parallel corpus

- **What we don't have:**

- Datasets for training helpful chatbots
- Instructions datasets to align models with human needs

Prompting, before instructions

Tatar - Detected English Spanish French

Бүгенге очрашуның темасы:
Зур Шартлаудан соң иң борынгы дәлиллэнгән галактика буларак JADES-GS-z13-0 билгеләнде. Жңир орбитасындагы Джеймс Уэбб инфракызыл нурланыш телескопы ярдәмендәге ачыш космологик кызыл тайпылманы үлчәүгә нигезләнде.

Менә бүгенге очрашуның транскрипциясе:
– Сәлам! Сез яңалыклар турында ишеттегезме?
– К. Кызык нәрсә укыдыгызмы?
– Әйе: 13 миллиард яктылык елы ераклыгында урнашкан иң борынгы галактиканың – JADES GS-Z13-0 булуын ачыкладылар.
– Ничек?
– Галактиканың яше 13,2 миллиард ел дип исәпләнә. Бу Галәмнең яше якынча 13,8 миллиард ел дип санала.
– Димәк, галактиканың яше Галәмнең яше белән бер чама.
– Әйе.
– Ә бу галактика ничек табылган?
– Ул Жңир орбитасында булган Джеймс Вебб галәми телескопы ярдәмендә ачылган.

745 / 5,000

English Spanish Arabic

The topic of today's meeting is:
JADES-GS-z13-0 has been identified as the oldest known galaxy since the Big Bang. The discovery by the James Webb Infrared Telescope in Earth orbit was based on measurements of cosmological redshift.

Here is a transcript of today's meeting:
- Hello! Have you heard about the news?
- No. Did you read something interesting?
- Yes: JADES GS-Z13-0, the oldest galaxy located 13 billion light-years away, has been discovered.
- How?
- The age of the galaxy is estimated to be 13.2 billion years. The age of this Universe is estimated to be about 13.8 billion years.
- Therefore, the age of the galaxy is about the same as the age of the Universe.
- Yes.
- How was this galaxy found?
- It was discovered by the James Webb Space Telescope in Earth orbit.

Not every task needs instruction- prompts.

We can achieve a lot using the "old" techniques people used prior to GPT-3.

Here, we use the GPT model to convert text from one style into another.

This helps with the low-resource issue!

Zero-Shot Translation

- By creating "Hydra" models which can work with both the new and the old tokenizer, we can use the models to translate from English to Tatar, beating Google Translate!

Әлбәттә, караган период өчен караучыга һел саен 1,8 пенсия коэффициенты күләмендә иминият стажы языла.

Of course, for the period covered, the carer is insured every month in the amount of 1.8 pension ratio.

45 chrF (Google):

Әлбәттә, капланган чор өчен карьера ай саен 1,8 пенсия күләмендә иминләштерелә.

60 chrF (Hydra):

Әлбәттә, капланган чор өчен ай саен 1,8 пенсия коэффициенты күләмендә иминиятләштерелә.

Despite having never seen any parallel sentence!

Zero-Shot Translation

- We can also translate from many languages, zero-shot, without using a pivot language!

```
[9]: print(translate_english_text("Bad property debt exceeds reserves at largest US banks: Loan loss provisions have thin  
# At the biggest U.S. banks, bad loans are outpacing reserves: Loan loss reserves are shrinking, even as watchdogs s
```

АКШның иң зур банкларында начар милек кредиты запас фондтан артып китә: кредит югалтулары өчен резервлар кими, хәтта күзәтчелек органнары коммерция күчемсез милек базарындагы хәвефләрне ассызыклайлар.

```
[10]: print(translate_english_text("""Qu'avez-vous fait auprès des citoyens ? Zéro" : tension lors du conseil communal de  
# "What did he do to the people?" – "Zero": tension in the Charleroi city council over the issue of pollution by met
```

«Халык белән нәрсә эшләгән?» – «Ноль»: Шарлеруа шәһәр шурасында металл кисүчеләр пычрату мәсьәләсе буенча килеренкелек.

```
[11]: print(translate_english_text("""Aan het station in Leuven is een vrouw van 27 gegrepen door een bus. Ze raakte zwaar  
# A 27-year-old woman is covered in a bus at Leuven station. He suffered serious injuries and was resuscitated at the
```

Левен станциясендә бер хатын-кыз 27 яшендә автобуста каплана. Ул авыр тән жәрәхәтләре алды һәм урында реаниматизация ләнде. Хатын-кыз әле больницага озатылды, әмма анда жәрәхәтләрәннән вафат булды. Ничек һәлакәт була алган, әле тикшерелә.

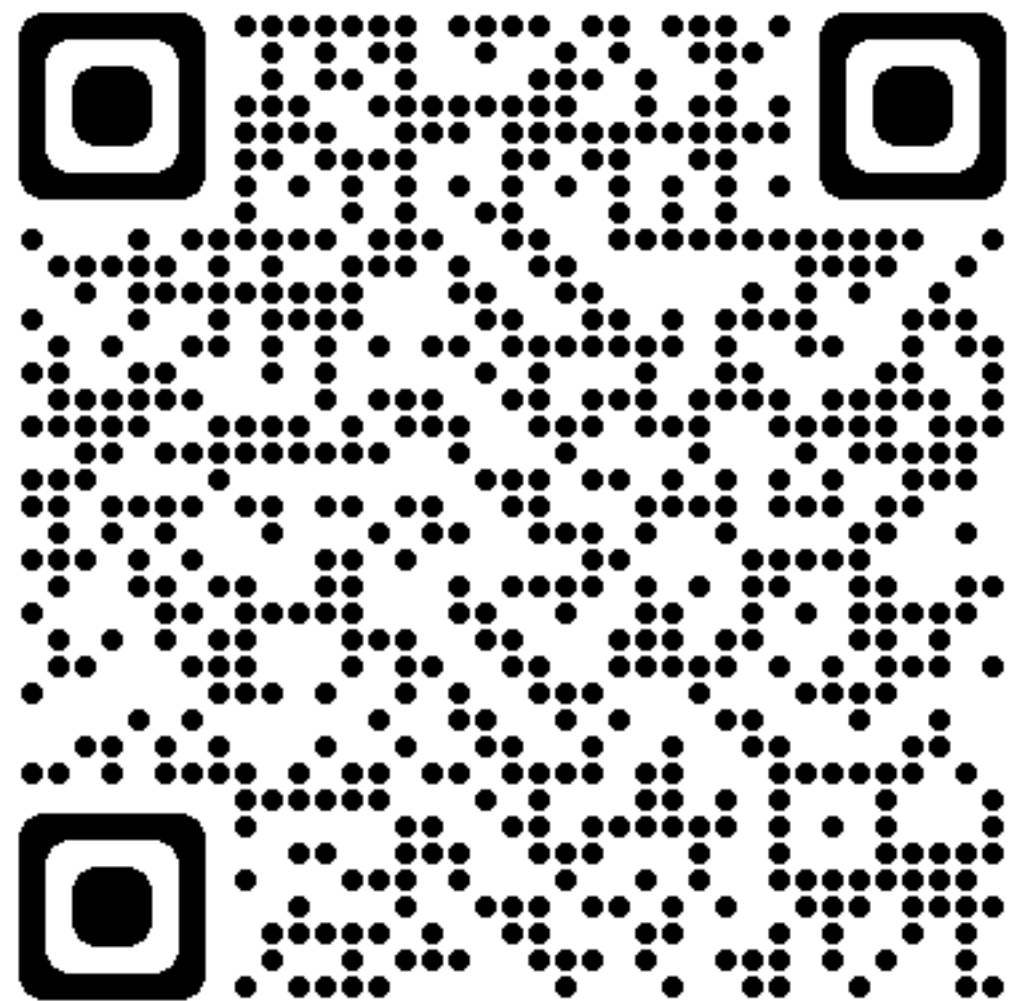
And the future!



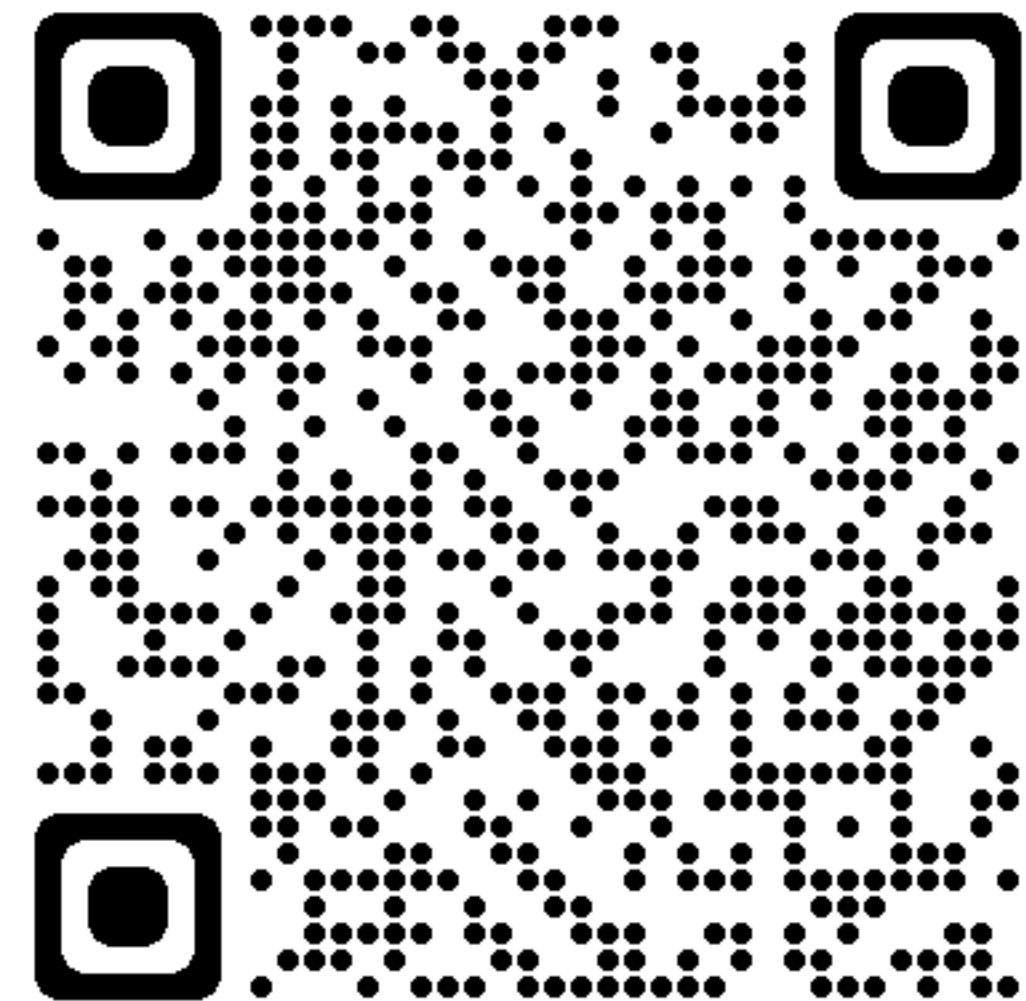
✨ **Tik-to-tok** ✨

Turn your monolingual LLMs into multilingual maven

Pieter Delobelle & François Remy
February 27, 2024



pieter.ai/tik-to-tok



twitter.com/fremycompany