

Dutch language models

Six years of tokenizers, data and RobBERTs

Dr. ing. Pieter Delobelle

Lead AI scientist @ Pleias

Guest professor @ KU Leuven



Who am I?

2026– Lead AI scientist at Pleias
2025– Guest professor at KU Leuven
2024–2025 LLM engineer at Aleph Alpha
2023 Apple

PhD and postdoc at KU Leuven's DTAI group

First author of RobBERT

Dutch BERT model, almost 5M downloads

Researching LLM training

tokenization, synthetic data, training, inference

Expert advisor, EU AI Act Code of Practice



EU AI Office's Network of Evaluators Workshop, April
2025

Dutch is a mid-resource language

- About 25M native speakers: the Netherlands, Flanders, Suriname

Dutch is a mid-resource language

- About 25M native speakers: the Netherlands, Flanders, Suriname
- In 2019: [just enough](#) Dutch web text to pretrain a BERT from scratch

Dutch is a mid-resource language

- About 25M native speakers: the Netherlands, Flanders, Suriname
- In 2019: [just enough](#) Dutch web text to pretrain a BERT from scratch
- English models are bigger and trained longer: no large Dutch BERT until 2023

Dutch is a mid-resource language

- About 25M native speakers: the Netherlands, Flanders, Suriname
- In 2019: [just enough](#) Dutch web text to pretrain a BERT from scratch
- English models are bigger and trained longer: no large Dutch BERT until 2023
- Multilingual models cover Dutch, but as one language among 100+

Dutch is a mid-resource language

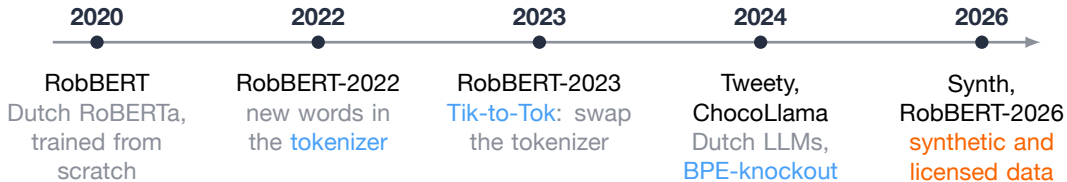
- About 25M native speakers: the Netherlands, Flanders, Suriname
- In 2019: **just enough** Dutch web text to pretrain a BERT from scratch
- English models are bigger and trained longer: no large Dutch BERT until 2023
- Multilingual models cover Dutch, but as one language among 100+

The question of this talk

What makes a language model **Dutch**? Its tokenizer? Its data? Both?

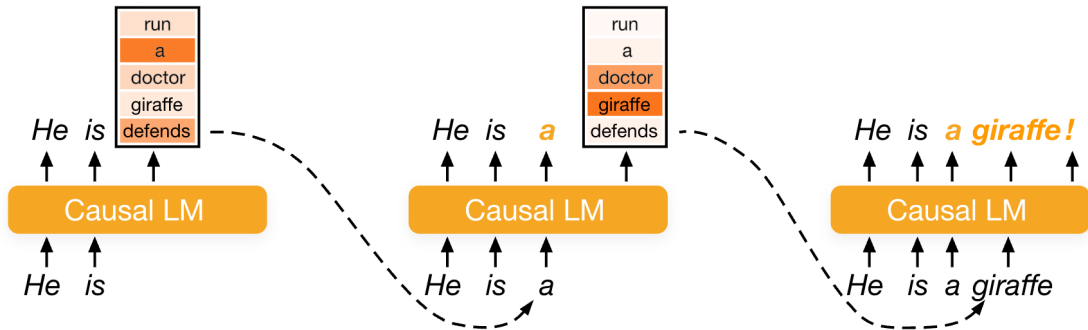
Six years of Dutch models

Three problems keep coming back: **tokenizers**, **data** and provenance



How language models read Dutch

Language models predict the next token



A **causal LM** (GPT, Llama) predicts the next token, one at a time.

A **masked LM** (BERT, RobBERT) fills in blanks, using context on both sides.

A masked LM fills in the blank

RobBERT on “Ik zie een <mask> in mijn tuin.” (I see a <mask> in my garden.)

Prediction	P
lamp (lamp)	0.40
boom (tree)	0.15
camera	0.09
ster (star)	0.05
stip (dot)	0.01

- Pretraining: mask 15–20% of tokens and predict them

A masked LM fills in the blank

RobBERT on “Ik zie een <mask> in mijn tuin.” (I see a <mask> in my garden.)

Prediction	P
lamp (lamp)	0.40
boom (tree)	0.15
camera	0.09
ster (star)	0.05
stip (dot)	0.01

- Pretraining: mask 15–20% of tokens and predict them
- Then fine-tune on a small labelled set (sentiment, NER, ...)

A masked LM fills in the blank

RobBERT on “Ik zie een <mask> in mijn tuin.” (I see a <mask> in my garden.)

Prediction	P
lamp (lamp)	0.40
boom (tree)	0.15
camera	0.09
ster (star)	0.05
stip (dot)	0.01

- Pretraining: mask 15–20% of tokens and predict them
- Then fine-tune on a small labelled set (sentiment, NER, ...)
- **giraf** (giraffe) is not even in the top 10k

Text becomes tokens, tokens become vectors

No, I am not a giraffe.

No, I am not a giraffe.

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

314: _I



21223: affe

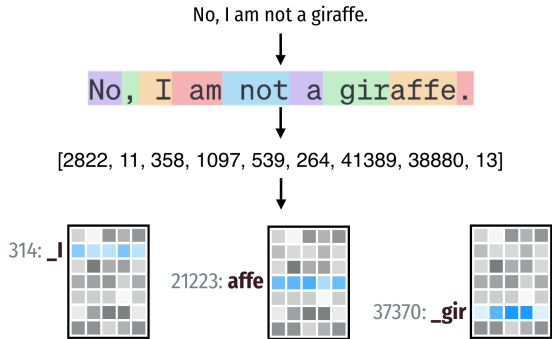


37370: _gir



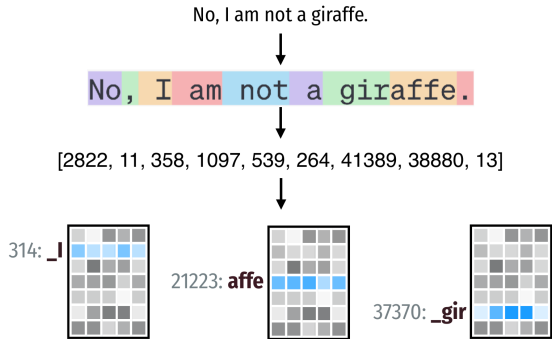
- The **tokenizer** splits text into subwords from a fixed vocabulary

Text becomes tokens, tokens become vectors



- The **tokenizer** splits text into subwords from a fixed vocabulary
- Each token id selects one row of the **embedding matrix**

Text becomes tokens, tokens become vectors



- The **tokenizer** splits text into subwords from a fixed vocabulary
- Each token id selects one row of the **embedding matrix**
- In RobBERT-base that matrix is about **30%** of all weights

Byte-pair encoding (BPE) merges frequent pairs

Ġ s t a n d a a r d t a r i e f

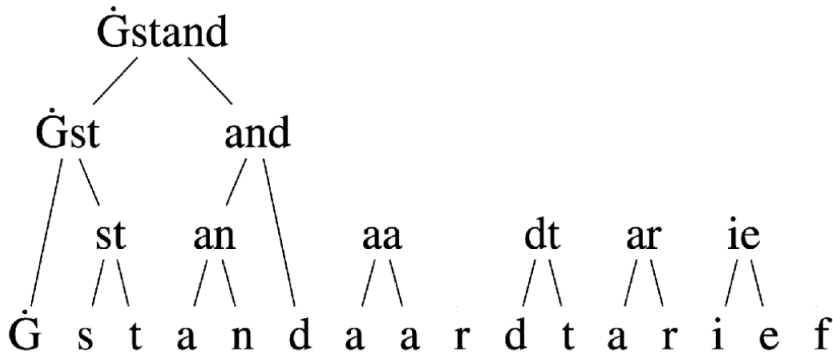
Start from bytes: Ġ marks the space before *standaardtarief* (standard rate).

Byte-pair encoding (BPE) merges frequent pairs

an
/ \
G s t a n d a a r d t a r i e f

Training: find the most frequent adjacent pair in the corpus and merge it.

Byte-pair encoding (BPE) merges frequent pairs



Repeat until the vocabulary is full. At inference, replay the merges in order.

Few non-English words are tokens

An English-centric tokenizer falls back to short, meaningless pieces

EN No, I am not a giraffe. That is an absurd thought. → fertility = 1.09

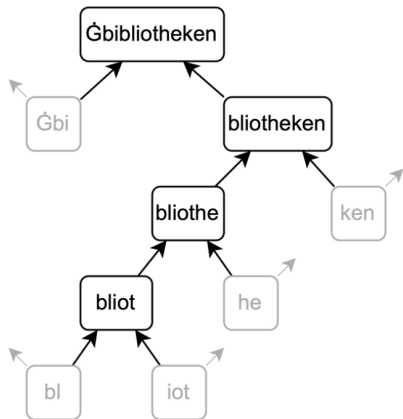
NL Nee, ik ben geen giraf. Dat is een absurde gedachte. → fertility = 1.55

DE Nein, ich bin keine Giraffe. Das ist ein absurder Gedanke. → fertility = 1.55

Fertility = tokens per word. 1.55 for Dutch means longer sequences, slower inference and less context.

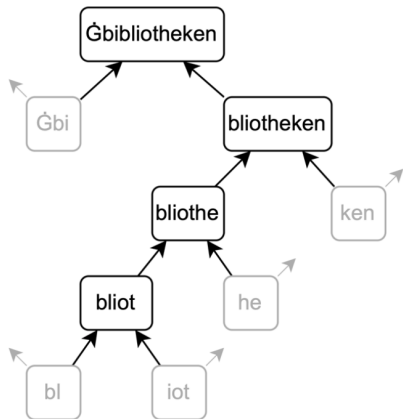
Chain effect: the vocabulary fills with unused tokens

- BPE only merges **two** existing tokens



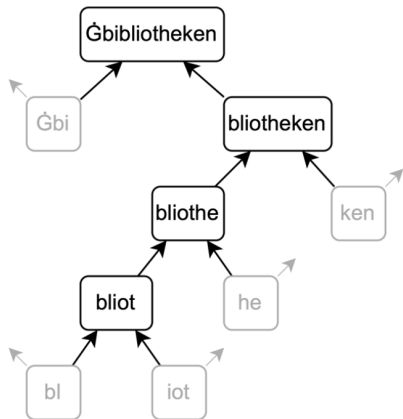
Chain effect: the vocabulary fills with unused tokens

- BPE only merges **two** existing tokens
- So Ġbibliotheken (libraries) needs bliot, bliothe, ... first



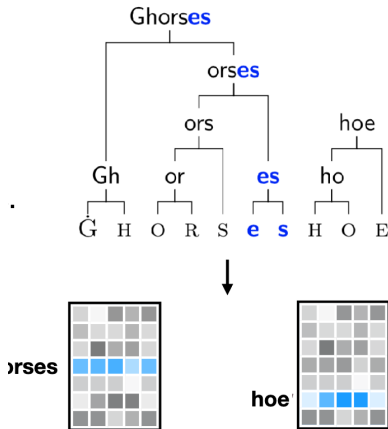
Chain effect: the vocabulary fills with unused tokens

- BPE only merges **two** existing tokens
- So Ġbibliotheken (libraries) needs bliot, bliothe, ... first
- Those intermediate tokens are rarely used on their own, but each takes an embedding row



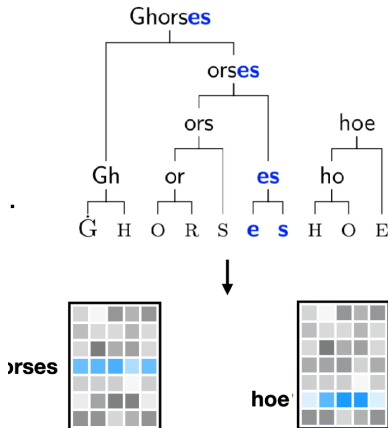
... and morpheme boundaries are not respected

- Tokenization happens *eagerly*: e+s merges early because it is frequent in English



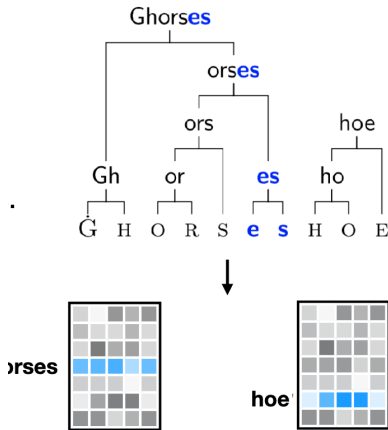
... and morpheme boundaries are not respected

- Tokenization happens *eagerly*: e+s merges early because it is frequent in English
- *horseshoe* becomes Ghorses + hoe, not *horse* + *shoe*



... and morpheme boundaries are not respected

- Tokenization happens *eagerly*: e+s merges early because it is frequent in English
- *horseshoe* becomes Ghorses + hoe, not *horse* + *shoe*
- Representations depend on tokens: a problem for **compounding** languages like Dutch



Why not give every word its own token?

- The vocabulary size is limited
 - Not every word in every language can be a token
 - Every token costs a row: $V \times h$ weights



Why not give every word its own token?

- The vocabulary size is limited
 - Not every word in every language can be a token
 - Every token costs a row: $V \times h$ weights
- But we want our target language represented well
 - More meaningful representations per token
 - Lower fertility = faster inference, more context



Why not give every word its own token?

- The vocabulary size is limited
 - Not every word in every language can be a token
 - Every token costs a row: $V \times h$ weights
- But we want our target language represented well
 - More meaningful representations per token
 - Lower fertility = faster inference, more context

→ **language-specific (or domain-specific) models**



RobBERT: a Dutch RoBERTa (2020)

RobBERT: a Dutch RoBERTa, trained from scratch

Delobelle, Winters & Berendt (2020), Findings of EMNLP

- RoBERTa-base architecture, masked language modelling



RobBERT: a Dutch RoBERTa, trained from scratch

Delobelle, Winters & Berendt (2020), Findings of EMNLP

- RoBERTa-base architecture, masked language modelling
- Dutch section of OSCAR: 39 GB, 6.6B words of web crawl



RobBERT: a Dutch RoBERTa, trained from scratch

Delobelle, Winters & Berendt (2020), Findings of EMNLP

- RoBERTa-base architecture, masked language modelling
- Dutch section of OSCAR: 39 GB, 6.6B words of web crawl
- v1 reused RoBERTa's **English** tokenizer; v2 got its own **Dutch** BPE tokenizer (40k tokens)



RobBERT: a Dutch RoBERTa, trained from scratch

Delobelle, Winters & Berendt (2020), Findings of EMNLP

- RoBERTa-base architecture, masked language modelling
- Dutch section of OSCAR: 39 GB, 6.6B words of web crawl
- v1 reused RoBERTa's **English** tokenizer; v2 got its own **Dutch** BPE tokenizer (40k tokens)
- *rob* is Dutch for seal, so the logo is a seal dressed as Bert



RobBERT became the default Dutch encoder

Used for

- Sentiment of book reviews: 94% accuracy
- *die* vs. *dat*: 98% correct
- NER, POS tagging, coreference
- Emotions, humour, cyberbullying
- Fixing *dt*-spelling mistakes

Still used today

- Almost [5M downloads](#)
- A family: RobBERTje (distilled), RobBERT-2022, -2023, -2026



Language keeps moving (2022)

2020 happened: corona is no longer just a beer

New tokens from a BPE tokenizer trained on OSCAR-2022

- RobBERT was trained in 2019: it never saw COVID-19

2020 happened: corona is no longer just a beer

New tokens from a BPE tokenizer trained on OSCAR-2022

- RobBERT was trained in 2019: it never saw COVID-19
- Train a new tokenizer on 2022 data and diff the vocabularies

2020 happened: corona is no longer just a beer

New tokens from a BPE tokenizer trained on OSCAR-2022

- RobBERT was trained in 2019: it never saw COVID-19
- Train a new tokenizer on 2022 data and diff the vocabularies
- The new tokens show what society started talking about

coronamaatregelen (corona measures)

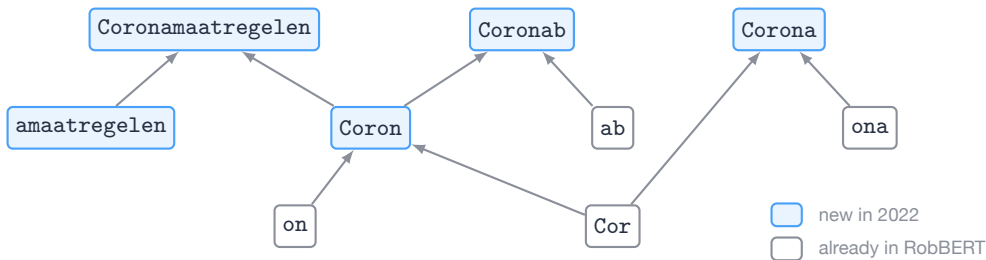
coronacrisis

coronabesmettingen (infections)

omikronvariant

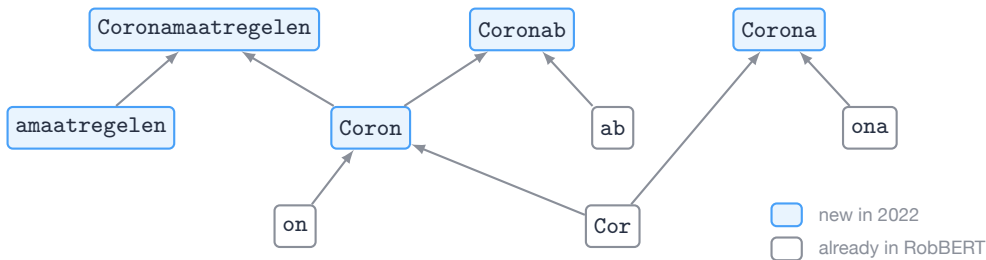
TikTok Biden Bitcoin

Adding a token means adding its whole merge tree



- Keep every merge that produces a token in **New – Old**: 2,774 tokens, vocabulary 40k
→ 42,774

Adding a token means adding its whole merge tree



- Keep every merge that produces a token in **New – Old**: 2,774 tokens, vocabulary 40k → 42,774
- coronamaatregelen: 5 tokens with the old tokenizer, 1 with the new one

RobBERT-2022: keep the weights, extend the vocabulary

Delobelle, Winters & Berendt (2022)

- Keep the 40k old embeddings, only learn the new rows

VaccinChat	Acc	F_1
BERTje	74.7	–
RobBERT v2	74.9	77.2
RobBERT-2022	76.3	79.3

Domain-adapted models reach 77.9 (Buhmann et al., 2022)

RobBERT-2022: keep the weights, extend the vocabulary

Delobelle, Winters & Berendt (2022)

- Keep the 40k old embeddings, only learn the new rows
- Continue pretraining on OSCAR-2022, whole documents instead of shuffled lines

VaccinChat	Acc	F_1
BERTje	74.7	–
RobBERT v2	74.9	77.2
RobBERT-2022	76.3	79.3

Domain-adapted models reach 77.9 (Buhmann et al., 2022)

RobBERT-2022: keep the weights, extend the vocabulary

Delobelle, Winters & Berendt (2022)

- Keep the 40k old embeddings, only learn the new rows
- Continue pretraining on OSCAR-2022, whole documents instead of shuffled lines
- Better on new topics, but on tasks about pre-2019 text the original can still win

VaccinChat	Acc	F_1
BERTje	74.7	–
RobBERT v2	74.9	77.2
RobBERT-2022	76.3	79.3

Domain-adapted models reach 77.9 (Buhmann et al., 2022)

Swapping tokenizers (2023–2024)

Pretraining is expensive, but worth it



One book

40-50k tokens



One bookshelf

1.6M - 2.5M tokens

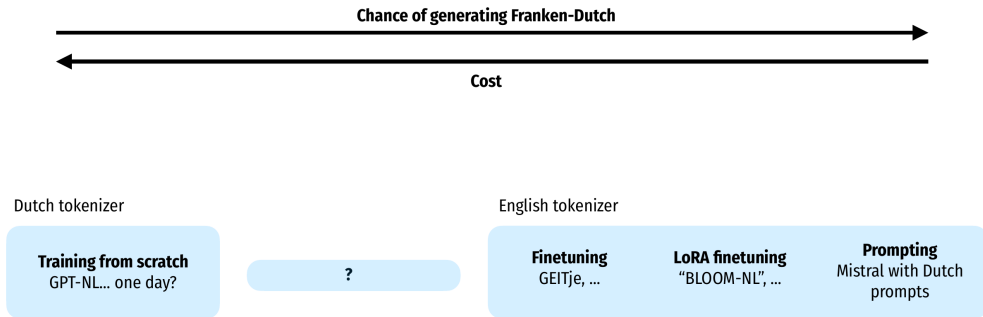


One LLM training set

2.5T - 6T tokens
~2 500 000 bookshelves

A clean, new tokenizer means a new embedding matrix. Do we have to throw away all GPU hours and start again?

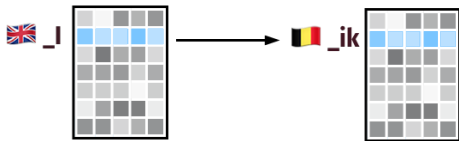
The options for a new language



An English tokenizer is cheap, but the model drifts into **Franken-Dutch**.
We want the "?": a Dutch tokenizer, without training from scratch.

Tik-to-Tok: translate the embeddings, token by token

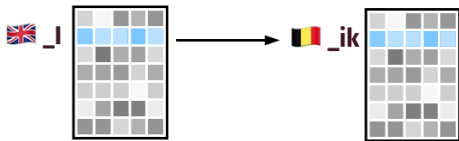
Remy, Delobelle, Berendt, Demuyne & Demeester (2023)



- Dutch `_ik` should start from the embedding of English `_I`

Tik-to-Tok: translate the embeddings, token by token

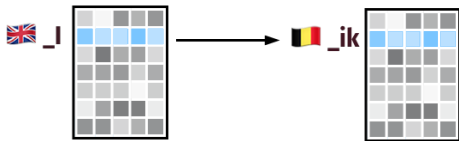
Remy, Delobelle, Berendt, Demuyne & Demeester (2023)



- Dutch `_ik` should start from the embedding of English `_I`
- Keep all transformer layers; only the embedding matrix is rebuilt

Tik-to-Tok: translate the embeddings, token by token

Remy, Delobelle, Berendt, Demuyne & Demeester (2023)



- Dutch `_ik` should start from the embedding of English `_I`
- Keep all transformer layers; only the embedding matrix is rebuilt
- Then continue pretraining on Dutch, for a fraction of the original cost

How do we map tokens that are not words?

WORD TRANSLATION DICTIONARY



1. A word translation dictionary aligns whole words

(3) FAST-TEXT WITH SKIPGRAM OBJECTIVE

$$f(\text{🇳🇱} \text{ VRIJ } + \dots + \text{ HEID } \text{🇳🇱})$$

SHOULD BE EQUAL TO

$$f(\text{🇺🇸} \text{ LIBER } + \dots + \text{ TY } \text{🇺🇸})$$

(4) TOKEN EMBEDDING + SOURCE NEIGHBORS RETRIEVAL

A diagram showing token embedding retrieval. It features the equation $f(\text{HEID } \text{🇳🇱}) \simeq f(\text{TY } \text{🇺🇸}), f(\text{RTY } \text{🇺🇸}), \dots$. The word 'HEID' is in a blue block, and 'TY' and 'RTY' are in green blocks. A green arrow points to the 'HEID' block from the left.

How do we map tokens that are not words?

WORD TRANSLATION DICTIONARY



(3) FAST-TEXT WITH SKIPGRAM OBJECTIVE

$$f(\text{Dutch flag}) \text{ VRIJ } + \dots + \text{ HEID } (\text{Dutch flag})$$

SHOULD BE EQUAL TO

$$f(\text{US flag}) \text{ LIBER } + \dots + \text{ TY } (\text{US flag})$$

(4) TOKEN EMBEDDING + SOURCE NEIGHBORS RETRIEVAL

A diagram showing token embedding retrieval. It features the equation $f(\text{HEID } \text{Dutch flag}) \simeq f(\text{TY } \text{US flag}), f(\text{RTY } \text{US flag}), \dots$. The word 'HEID' is in a blue block, and 'TY' and 'RTY' are in green blocks. A green arrow points to the left side of the equation.

1. A word translation dictionary aligns whole words
2. FastText with character n -grams embeds both languages in one space

How do we map tokens that are not words?

WORD TRANSLATION DICTIONARY



(3) FAST-TEXT WITH SKIPGRAM OBJECTIVE

$$f(\text{Dutch flag}) \text{ VRIJ } + \dots + \text{ HEID } (\text{Dutch flag})$$

SHOULD BE EQUAL TO

$$f(\text{US flag}) \text{ LIBER } + \dots + \text{ TY } (\text{US flag})$$

(4) TOKEN EMBEDDING + SOURCE NEIGHBORS RETRIEVAL

A diagram showing token embedding retrieval. It features the equation $f(\text{HEID } \text{Dutch flag}) \simeq f(\text{TY } \text{US flag}), f(\text{RTY } \text{US flag}), \dots$. The word 'HEID' is in a blue box, and 'TY' and 'RTY' are in green boxes. A green arrow points to the equation.

1. A word translation dictionary aligns whole words
2. FastText with character n -grams embeds both languages in one space
3. Embed every target token, retrieve its nearest source tokens and combine their embeddings

RobBERT-2023 tops the DUMB benchmark

RoBERTa-large converted to Dutch with Tik-to-Tok, on a cleaned OSCAR-2023

Model	Avg. RER
RobBERT-2023 large	18.6
DeBERTa-v3 large (EN)	15.7
XLNet-R large	14.4
RobBERT-2023 base	4.5
RobBERT-2022 base	3.6
RobBERT v2 base	1.6
BERTje base	0.0
RobBERT v1 base	-16.3

- RER: relative error reduction vs. BERTje, averaged over 9 tasks

DUMB: de Vries, Wieling & Nissim (2023)

RobBERT-2023 tops the DUMB benchmark

RoBERTa-large converted to Dutch with Tik-to-Tok, on a cleaned OSCAR-2023

Model	Avg. RER
RobBERT-2023 large	18.6
DeBERTa-v3 large (EN)	15.7
XLM-R large	14.4
RobBERT-2023 base	4.5
RobBERT-2022 base	3.6
RobBERT v2 base	1.6
BERTje base	0.0
RobBERT v1 base	-16.3

- RER: relative error reduction vs. BERTje, averaged over 9 tasks
- A converted English model beats every Dutch model trained from scratch

DUMB: de Vries, Wieling & Nissim (2023)

RobBERT-2023 tops the DUMB benchmark

RoBERTa-large converted to Dutch with Tik-to-Tok, on a cleaned OSCAR-2023

Model	Avg. RER
RobBERT-2023 large	18.6
DeBERTa-v3 large (EN)	15.7
XLM-R large	14.4
RobBERT-2023 base	4.5
RobBERT-2022 base	3.6
RobBERT v2 base	1.6
BERTje base	0.0
RobBERT v1 base	-16.3

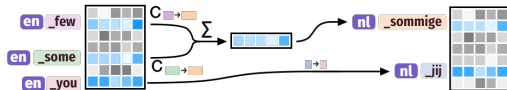
- RER: relative error reduction vs. BERTje, averaged over 9 tasks
- A converted English model beats every Dutch model trained from scratch
- RobBERT v1 vs. v2: an English vs. a Dutch tokenizer, -16.3 vs. +1.6

DUMB: de Vries, Wieling & Nissim (2023)

Trans-tokenization: the same idea for LLMs

Remy et al. (2024), COLM

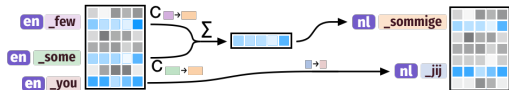
- Align tokens with FastAlign on a parallel corpus



Trans-tokenization: the same idea for LLMs

Remy et al. (2024), COLM

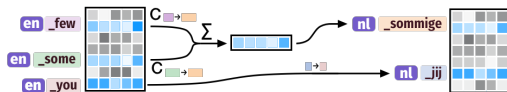
- Align tokens with FastAlign on a parallel corpus
- A Dutch token's embedding is the **weighted average** of the English tokens it aligns to



Trans-tokenization: the same idea for LLMs

Remy et al. (2024), COLM

- Align tokens with FastAlign on a parallel corpus
- A Dutch token's embedding is the **weighted average** of the English tokens it aligns to
- Tweety-7B-Dutch: Mistral-7B with a Dutch tokenizer, 8.5B tokens of further training



A Dutch tokenizer needs a third fewer tokens

gpt-neo-1.3b-dutch, **tweety-7b-dutch**

Tokens: 15

Nee , ik ben helemaal geen gir af ! Dat is een absurde gedachte .

mistral-7b, geitje-7b

Tokens: 23

N ee , ik ben he le ma al ge en gir af ! Dat is een abs ur de ged ach
te .

23 → 15 tokens here; about 33% fewer tokens overall, so faster and cheaper.

Hydra models: one body, a head per language

Trans-tokenization also works for low-resource languages, like Tatar

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):

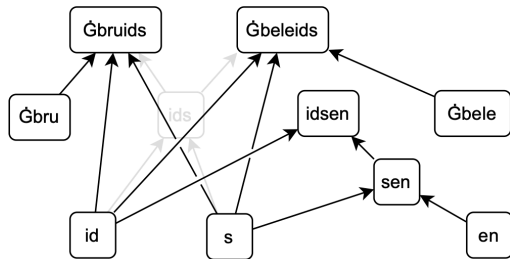
Әлбәттә, капланган чор өчен ай саен 1,8 пенсия коэффициенты күләмендә иминиятләштерелә.

Swap the input and output heads and the model translates English → Tatar without training on translation. Here: 60 chrF vs. 45 for Google Translate.

BPE-knockout: remove the merges that hurt

Bauwens & Delobelle (2024), NAACL

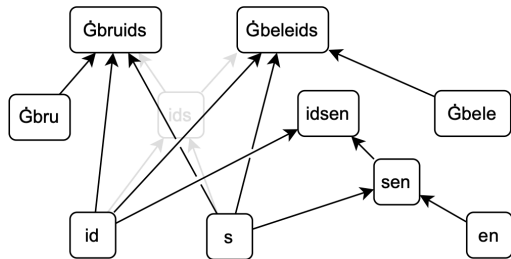
- Compare BPE with a morphological segmentation (*bruid* | *s*, brides)



BPE-knockout: remove the merges that hurt

Bauwens & Delobelle (2024), NAACL

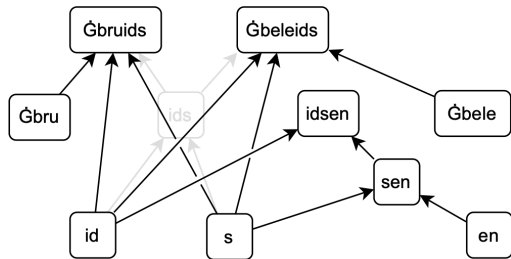
- Compare BPE with a morphological segmentation (*bruid* | *s*, brides)
- Blame each lost boundary on exactly one merge



BPE-knockout: remove the merges that hurt

Bauwens & Delobelle (2024), NAACL

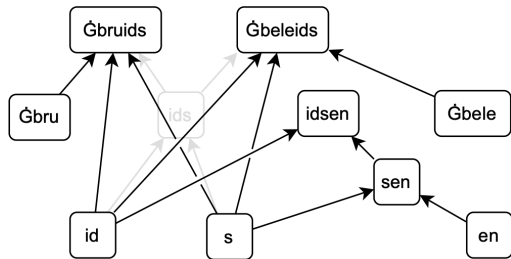
- Compare BPE with a morphological segmentation (*bruid* | *s*, brides)
- Blame each lost boundary on exactly one merge
- Knock out merges that are blamed in $\geq \frac{1}{2}$ of their uses



BPE-knockout: remove the merges that hurt

Bauwens & Delobelle (2024), NAACL

- Compare BPE with a morphological segmentation (*bruid* | *s*, brides)
- Blame each lost boundary on exactly one merge
- Knock out merges that are blamed in $\geq \frac{1}{2}$ of their uses
- Descendants survive: $(xy, z) \rightarrow (x, y, z)$



Dutch LLMs: ChocoLlama (2024)

ChocoLlama: more effort to curate the data

Meeus, Rathé, Remy, Delobelle, Decorte & Demeester (2024)

Source	GB	Tokens
OSCAR (Common Crawl)	93	28.6B
Open Subtitles	5	1.5B
Wikipedia	2.5	769M
Job descriptions (TechWolf)	1.5	462M
Staatsblad (Bizzy)	1.4	431M
Project Gutenberg	0.3	92M
Legislation (ML6)	0.2	62M

- 104 GB, 32B tokens in total



Computerwetenschappers bouwen Vlaams AI-model ChocoLlama

06 februari 2025 16:48

ChocoLlama: more effort to curate the data

Meeus, Rathé, Remy, Delobelle, Decorte & Demeester (2024)

Source	GB	Tokens
OSCAR (Common Crawl)	93	28.6B
Open Subtitles	5	1.5B
Wikipedia	2.5	769M
Job descriptions (TechWolf)	1.5	462M
Staatsblad (Bizzy)	1.4	431M
Project Gutenberg	0.3	92M
Legislation (ML6)	0.2	62M

- 104 GB, 32B tokens in total
- Partners shared domain data



Computerwetenschappers bouwen Vlaams AI-model ChocoLlama

06 februari 2025 16:48

ChocoLlama: more effort to curate the data

Meeus, Rathé, Remy, Delobelle, Decorte & Demeester (2024)

Source	GB	Tokens
OSCAR (Common Crawl)	93	28.6B
Open Subtitles	5	1.5B
Wikipedia	2.5	769M
Job descriptions (TechWolf)	1.5	462M
Staatsblad (Bizzy)	1.4	431M
Project Gutenberg	0.3	92M
Legislation (ML6)	0.2	62M

- 104 GB, 32B tokens in total
- Partners shared domain data
- Still **90% web crawl**



Computerwetenschappers bouwen Vlaams AI-model ChocoLlama

06 februari 2025 16:48

Llama-3 already spoke Dutch

Model	ARC	HellaSwag	MMLU	TruthfulQA	Avg.
Llama-3-ChocoLlama-instruct	0.48	0.66	0.49	0.49	0.53
llama-3-8B-robatch	0.44	0.64	0.46	0.48	0.51
llama-3-8B-instruct	0.47	0.59	0.47	0.52	0.51
llama-3-8B	0.44	0.64	0.47	0.45	0.5
Reynaerde-7B-Chat	0.44	0.62	0.39	0.52	0.49
Llama-3-ChocoLlama-base	0.45	0.64	0.44	0.44	0.49
zephyr-7b-beta	0.43	0.58	0.43	0.53	0.49
geitje-7b-ultra	0.40	0.66	0.36	0.49	0.48
ChocoLlama-2-7B-tokentrans-instruct	0.45	0.62	0.34	0.42	0.46
mistral-7b-v0.1	0.43	0.58	0.37	0.45	0.46
ChocoLlama-2-7B-tokentrans-base	0.42	0.61	0.32	0.43	0.45
ChocoLlama-2-7B-instruct	0.36	0.57	0.33	0.45	**0.43
ChocoLlama-2-7B-base	0.35	0.56	0.31	0.43	0.41
llama-2-7b-chat-hf	0.36	0.49	0.33	0.44	0.41
llama-2-7b-hf	0.36	0.51	0.32	0.41	0.40

- Llama-2 adapted: small gains

Llama-3 already spoke Dutch

Model	ARC	HellaSwag	MMLU	TruthfulQA	Avg.
Llama-3-ChocoLlama-instruct	0.48	0.66	0.49	0.49	0.53
llama-3-8B-robatch	0.44	0.64	0.46	0.48	0.51
llama-3-8B-instruct	0.47	0.59	0.47	0.52	0.51
llama-3-8B	0.44	0.64	0.47	0.45	0.5
Reynaerde-7B-Chat	0.44	0.62	0.39	0.52	0.49
Llama-3-ChocoLlama-base	0.45	0.64	0.44	0.44	0.49
zephyr-7b-beta	0.43	0.58	0.43	0.53	0.49
geitje-7b-ultra	0.40	0.66	0.36	0.49	0.48
ChocoLlama-2-7B-tokentrans-instruct	0.45	0.62	0.34	0.42	0.46
mistral-7b-v0.1	0.43	0.58	0.37	0.45	0.46
ChocoLlama-2-7B-tokentrans-base	0.42	0.61	0.32	0.43	0.45
ChocoLlama-2-7B-instruct	0.36	0.57	0.33	0.45	**0.43
ChocoLlama-2-7B-base	0.35	0.56	0.31	0.43	0.41
llama-2-7b-chat-hf	0.36	0.49	0.33	0.44	0.41
llama-2-7b-hf	0.36	0.51	0.32	0.41	0.40

- Llama-2 adapted: small gains
- Llama-2 with a Dutch tokenizer (trans-tokenized) beats Llama-2 with its own: 0.46 vs. 0.43

Llama-3 already spoke Dutch

Model	ARC	HellaSwag	MMLU	TruthfulQA	Avg.
Llama-3-ChocoLlama-instruct	0.48	0.66	0.49	0.49	0.53
llama-3-8B-robatch	0.44	0.64	0.46	0.48	0.51
llama-3-8B-instruct	0.47	0.59	0.47	0.52	0.51
llama-3-8B	0.44	0.64	0.47	0.45	0.5
Reynaerde-7B-Chat	0.44	0.62	0.39	0.52	0.49
Llama-3-ChocoLlama-base	0.45	0.64	0.44	0.44	0.49
zephyr-7b-beta	0.43	0.58	0.43	0.53	0.49
geitje-7b-ultra	0.40	0.66	0.36	0.49	0.48
ChocoLlama-2-7B-tokentrans-instruct	0.45	0.62	0.34	0.42	0.46
mistral-7b-v0.1	0.43	0.58	0.37	0.45	0.46
ChocoLlama-2-7B-tokentrans-base	0.42	0.61	0.32	0.43	0.45
ChocoLlama-2-7B-instruct	0.36	0.57	0.33	0.45	**0.43
ChocoLlama-2-7B-base	0.35	0.56	0.31	0.43	0.41
llama-2-7b-chat-hf	0.36	0.49	0.33	0.44	0.41
llama-2-7b-hf	0.36	0.51	0.32	0.41	0.40

- Llama-2 adapted: small gains
- Llama-2 with a Dutch tokenizer (trans-tokenized) beats Llama-2 with its own: 0.46 vs. 0.43
- Llama-3 base already scores higher than every adapted Llama-2

Llama-3 already spoke Dutch

Model	ARC	HellaSwag	MMLU	TruthfulQA	Avg.
Llama-3-ChocoLlama-instruct	0.48	0.66	0.49	0.49	0.53
llama-3-8B-robatch	0.44	0.64	0.46	0.48	0.51
llama-3-8B-instruct	0.47	0.59	0.47	0.52	0.51
llama-3-8B	0.44	0.64	0.47	0.45	0.5
Reynaerde-7B-Chat	0.44	0.62	0.39	0.52	0.49
Llama-3-ChocoLlama-base	0.45	0.64	0.44	0.44	0.49
zephyr-7b-beta	0.43	0.58	0.43	0.53	0.49
geitje-7b-ultra	0.40	0.66	0.36	0.49	0.48
ChocoLlama-2-7B-tokentrans-instruct	0.45	0.62	0.34	0.42	0.46
mistral-7b-v0.1	0.43	0.58	0.37	0.45	0.46
ChocoLlama-2-7B-tokentrans-base	0.42	0.61	0.32	0.43	0.45
ChocoLlama-2-7B-instruct	0.36	0.57	0.33	0.45	**0.43
ChocoLlama-2-7B-base	0.35	0.56	0.31	0.43	0.41
llama-2-7b-chat-hf	0.36	0.49	0.33	0.44	0.41
llama-2-7b-hf	0.36	0.51	0.32	0.41	0.40

- Llama-2 adapted: small gains
- Llama-2 with a Dutch tokenizer (trans-tokenized) beats Llama-2 with its own: 0.46 vs. 0.43
- Llama-3 base already scores higher than every adapted Llama-2
- Better multilingual bases shrink the gain of continued pretraining

Synthetic data: Synth (2026)

Web crawls are running out of road

- Little control over what the model sees

Web crawls are running out of road

- Little control over what the model sees
- Copyright and opt-outs: sites block crawlers, datasets shrink

Web crawls are running out of road

- Little control over what the model sees
- Copyright and opt-outs: sites block crawlers, datasets shrink
- Built for pretraining only: little explicit reasoning, no instructions

Web crawls are running out of road

- Little control over what the model sees
- Copyright and opt-outs: sites block crawlers, datasets shrink
- Built for pretraining only: little explicit reasoning, no instructions
- Frontier labs build their own **synthetic** datasets for mid- and post-training, but none are public

Web crawls are running out of road

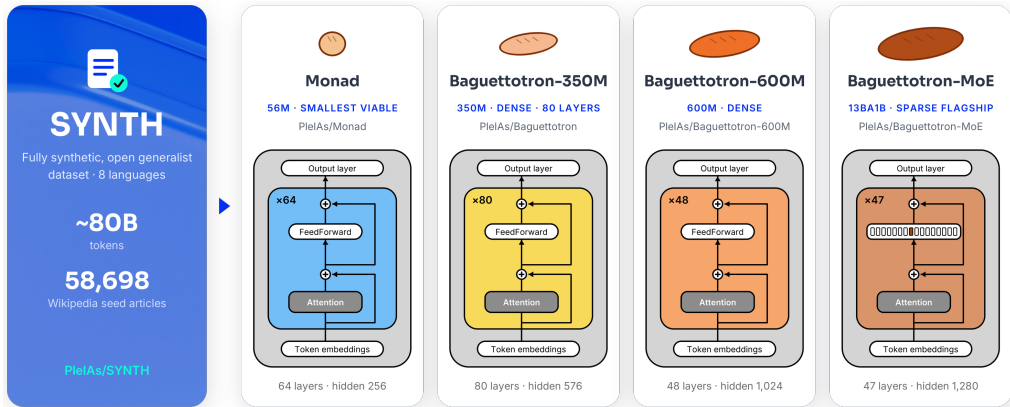
- Little control over what the model sees
- Copyright and opt-outs: sites block crawlers, datasets shrink
- Built for pretraining only: little explicit reasoning, no instructions
- Frontier labs build their own **synthetic** datasets for mid- and post-training, but none are public

Can we pretrain on synthetic data only, in one stage?

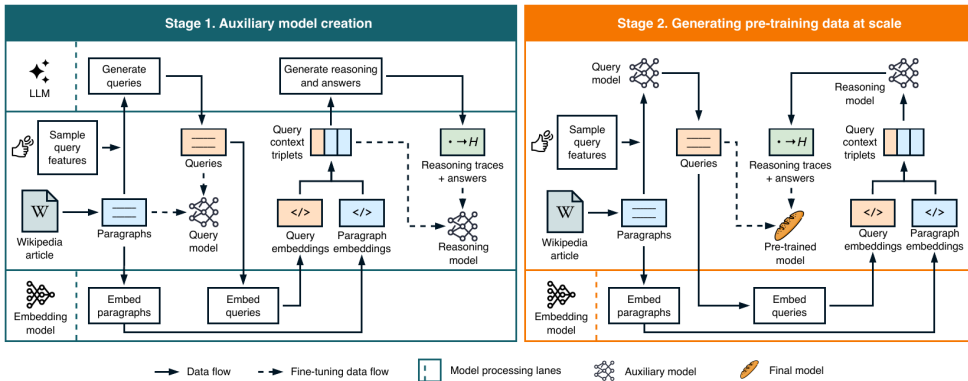
No separate SFT or RL stage: pre-, mid- and post-training in one corpus.

Synth: one synthetic dataset, a suite of models

Langlais, Delobelle et al. (2026), NeurIPS



Two stages: train small generators, then run them at scale



Generators are small fine-tunes (e.g. LoRA on Gemma-3-12B-base) on frontier-LLM outputs.

Seeds and constraints control what the model learns

Seeds

- 50k Wikipedia *vital articles*
- 8,698 articles on law, medicine, chemistry
- 3,727 Wikibooks pages (cooking)
- Every seed segment amplified 100×

Seeds and constraints control what the model learns

Seeds

- 50k Wikipedia *vital articles*
- 8,698 articles on law, medicine, chemistry
- 3,727 Wikibooks pages (cooking)
- Every seed segment amplified 100×

Constraints, sampled per query

- Six axes: type, complexity, user, result, language, style
- Forces diversity on axes we choose

Seeds and constraints control what the model learns

Seeds

- 50k Wikipedia *vital articles*
- 8,698 articles on law, medicine, chemistry
- 3,727 Wikibooks pages (cooking)
- Every seed segment amplified 100×

Constraints, sampled per query

- Six axes: type, complexity, user, result, language, style
- Forces diversity on axes we choose
- 20% target a refusal, correction or hedge

Reasoning traces in a stenographic syntax

- Not natural-language chain of thought, but compact markers:
 - logical: $\rightarrow$ $\circlearrowleft$ $\therefore$
 - epistemic: from $\bullet$ certain to $\circ$ uncertain
 - verification: $\square$ $\checkmark$

Reasoning traces in a stenographic syntax

- Not natural-language chain of thought, but compact markers:
 - logical: $\rightarrow$ $\circlearrowleft$ $\therefore$
 - epistemic: from $\bullet$ certain to $\circ$ uncertain
 - verification: $\square$ $\checkmark$
- Each marker is a **dedicated token**: the last ~ 50 entries of the 65,535-token vocabulary

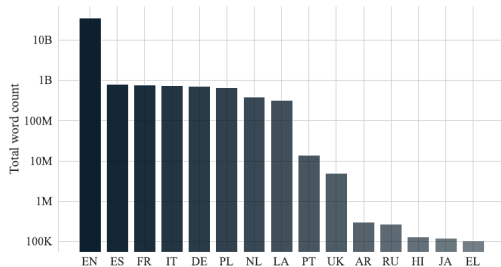
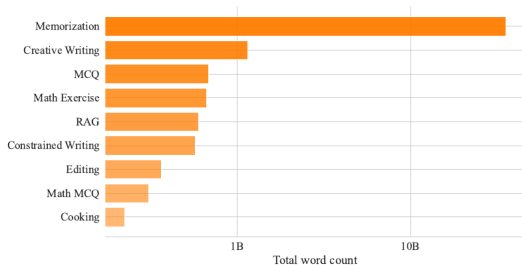
Reasoning traces in a stenographic syntax

- Not natural-language chain of thought, but compact markers:
 - logical: $\rightarrow$ $\circlearrowleft$ $\therefore$
 - epistemic: from $\bullet$ certain to $\circ$ uncertain
 - verification: $\square$ $\checkmark$
- Each marker is a **dedicated token**: the last ~ 50 entries of the 65,535-token vocabulary
- A small model fits a richer trace in a fixed context window

Reasoning traces in a stenographic syntax

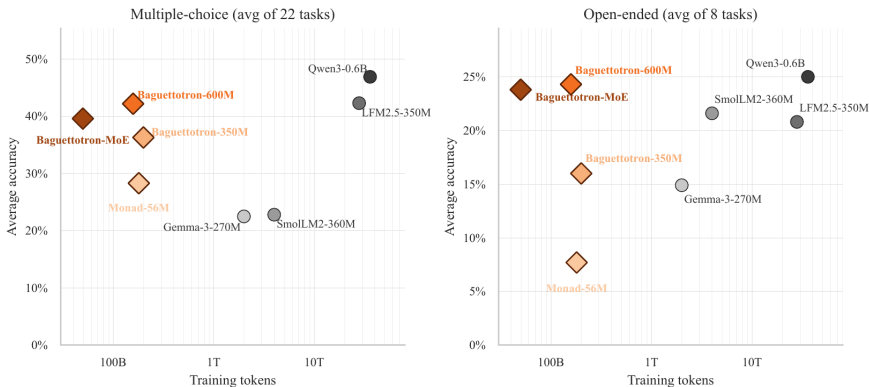
- Not natural-language chain of thought, but compact markers:
 - logical: $\rightarrow$ $\circlearrowleft$ $\therefore$
 - epistemic: from $\bullet$ certain to $\circ$ uncertain
 - verification: $\square$ $\checkmark$
- Each marker is a **dedicated token**: the last ~ 50 entries of the 65,535-token vocabulary
- A small model fits a richer trace in a fixed context window
- The markers become a usable confidence signal: uncertain traces state fewer and less precise facts

Tasks and languages



About 80B tokens in 8 languages; 20% is non-English, Dutch included. Reasoning traces stay in English.

Baguettotron sits on the token-efficiency frontier



Trails Qwen3-0.6B by 4.7 points (multiple choice) and 0.7 (open-ended), on **80–700×** fewer tokens.

Same model, same tokenizer, different data

A 600M model on Synth vs. on English web data

- Trained on FineWiki or FinePDFs-Edu, the model gives **no valid multiple-choice answers** at all

Same model, same tokenizer, different data

A 600M model on Synth vs. on English web data

- Trained on FineWiki or FinePDFs-Edu, the model gives **no valid multiple-choice answers** at all
- After post-training them on SmolTalk, Synth still leads by **16–17** points (multiple choice) and **11–14** (open-ended)

Same model, same tokenizer, different data

A 600M model on Synth vs. on English web data

- Trained on FineWiki or FinePDFs-Edu, the model gives **no valid multiple-choice answers** at all
- After post-training them on SmolTalk, Synth still leads by **16–17** points (multiple choice) and **11–14** (open-ended)
- Without reasoning traces: multiple choice unchanged, open-ended 24.3 → 22.0, TruthfulQA –9.1

Same model, same tokenizer, different data

A 600M model on Synth vs. on English web data

- Trained on FineWiki or FinePDFs-Edu, the model gives **no valid multiple-choice answers** at all
- After post-training them on SmolTalk, Synth still leads by **16–17** points (multiple choice) and **11–14** (open-ended)
- Without reasoning traces: multiple choice unchanged, open-ended 24.3 → 22.0, TruthfulQA –9.1
- Deliberately overtrained: 264 tokens per parameter (Chinchilla: 20), loss still going down

Grounded data gives more precise facts

“What do you know about {entity}?” on 500 seed entities, FActScore-style

Model	Tokens	Precision	Macro
Qwen3-0.6B	36T	65.7	31.6
Phi-4-mini-instruct (3.8B)	5T	77.4	29.5
Baguettotron-600M	158B	79.3	41.7
OLMoE-1B-7B-Instruct	5.1T	82.6	33.4
Baguettotron-MoE (13B/1B)	50B	81.7	46.3

On entities *outside* the seeds, Baguettotron-MoE abstains 67% of the time; OLMoE only 7%.

Pick the seeds, pick the domain

Telecom: 93M seed tokens (Wikipedia + 3GPP standards), amplified $10\times$

600M model	Base	+ telecom Synth
TeleQnA accuracy	41.6	56.7
FActScore on 3GPP	21.5	38.8
LLM judge (0–100)	27.8	67.8

The same works for a language: RobBERT-2026's synthetic Dutch is seeded from Dutch Wikipedia and the Belgian Staatsblad.

Read more about Synth



Paper

arxiv.org/pdf/2609.37891



Thread

[@pieterdelobelle on X](#)

RobBERT-2026

Now we can afford to be picky

- 2019: just enough Dutch, and per OSCAR, shuffling sentences sufficed for copyright

Now we can afford to be picky

- 2019: just enough Dutch, and per OSCAR, shuffling sentences sufficed for copyright
- 2026: the [EU AI Act](#) asks for data governance and transparency

Now we can afford to be picky

- 2019: just enough Dutch, and per OSCAR, shuffling sentences sufficed for copyright
- 2026: the [EU AI Act](#) asks for data governance and transparency
- Most Dutch models, every RobBERT included, were trained on web crawl

Now we can afford to be picky

- 2019: just enough Dutch, and per OSCAR, shuffling sentences sufficed for copyright
- 2026: the [EU AI Act](#) asks for data governance and transparency
- Most Dutch models, every RobBERT included, were trained on web crawl
- GPT-NL set out to build a compliant Dutch model, but never released it

Now we can afford to be picky

- 2019: just enough Dutch, and per OSCAR, shuffling sentences sufficed for copyright
- 2026: the [EU AI Act](#) asks for data governance and transparency
- Most Dutch models, every RobBERT included, were trained on web crawl
- GPT-NL set out to build a compliant Dutch model, but never released it

Can a Dutch model be competitive with only permissibly licensed or synthetic data?

RobBERT-2026

A Dutch encoder trained only on permissibly licensed and synthetic data



- **NeoBERT-base**: 275M parameters, trained from scratch

RobBERT-2026

A Dutch encoder trained only on permissibly licensed and synthetic data



- **NeoBERT-base**: 275M parameters, trained from scratch
- **~27B tokens**, permissibly licensed or synthetic

RobBERT-2026

A Dutch encoder trained only on permissibly licensed and synthetic data



- **NeoBERT-base**: 275M parameters, trained from scratch
- **~27B tokens**, permissibly licensed or synthetic
- **No web crawl**: no OSCAR, no CommonCrawl, no FineWeb

RobBERT-2026

A Dutch encoder trained only on permissibly licensed and synthetic data



- **NeoBERT-base**: 275M parameters, trained from scratch
- **~27B tokens**, permissibly licensed or synthetic
- **No web crawl**: no OSCAR, no CommonCrawl, no FineWeb
- As far as we know, the first released Dutch model that is explicitly [EU AI Act compliant](#)

RobBERT-2026

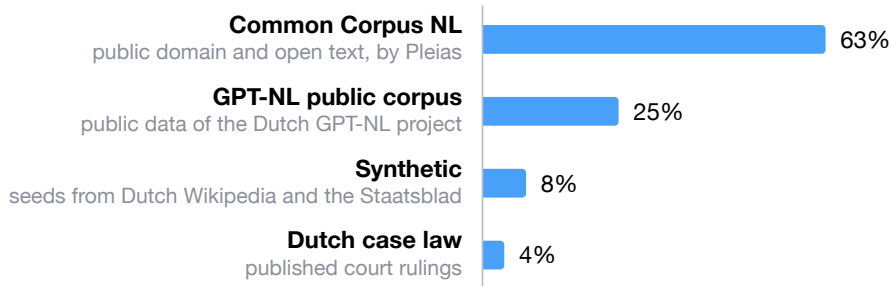
A Dutch encoder trained only on permissibly licensed and synthetic data



- **NeoBERT-base**: 275M parameters, trained from scratch
- **~27B tokens**, permissibly licensed or synthetic
- **No web crawl**: no OSCAR, no CommonCrawl, no FineWeb
- As far as we know, the first released Dutch model that is explicitly [EU AI Act compliant](#)
- **MIT licence**, plus a CC-BY-SA retrieval model

No web crawl lineage, of any kind

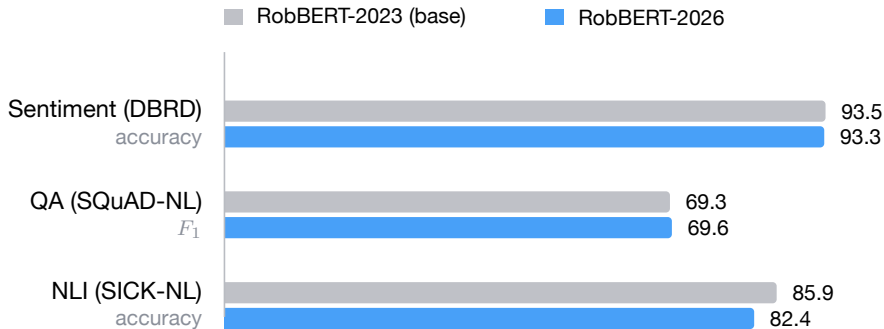
Share of the $\sim 27\text{B}$ pretraining tokens per source



No OSCAR or CommonCrawl, and no crawl in disguise: no CC-licensed crawl subset, no translated FineWeb-Edu.

Clean data costs little

DUMB test split, the same fine-tuning recipe for both models



Level on sentiment and question answering; natural language inference drops 3.5 points.

In context: mid-field, with a weak spot

Published DUMB scores for Dutch base-size encoders

Model	QA F_1	SA	NLI
Tik-to-Tok base	75.6	93.6	86.6
RobBERT v2 base	71.0	93.2	84.6
BERTje base	70.3	93.3	85.2
RobBERT-2026	69.6	93.3	82.4
RobBERT-2023 base	68.6	93.3	86.3

- Level on sentiment, mid-field on QA

In context: mid-field, with a weak spot

Published DUMB scores for Dutch base-size encoders

Model	QA F_1	SA	NLI
Tik-to-Tok base	75.6	93.6	86.6
RobBERT v2 base	71.0	93.2	84.6
BERTje base	70.3	93.3	85.2
RobBERT-2026	69.6	93.3	82.4
RobBERT-2023 base	68.6	93.3	86.3

- Level on sentiment, mid-field on QA
- NLI is the **weak spot**: below every listed model

In context: mid-field, with a weak spot

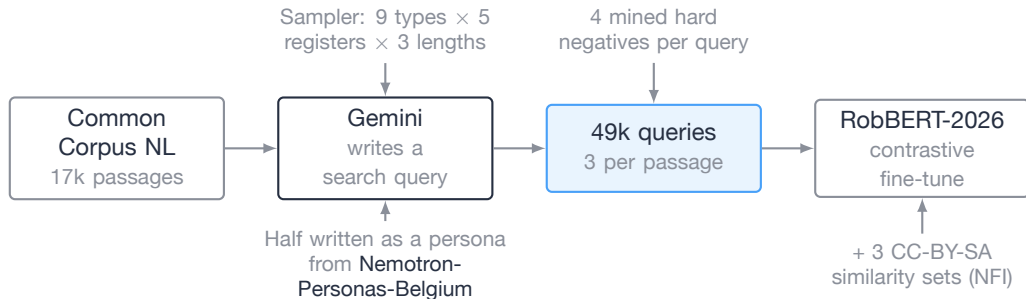
Published DUMB scores for Dutch base-size encoders

Model	QA F_1	SA	NLI
Tik-to-Tok base	75.6	93.6	86.6
RobBERT v2 base	71.0	93.2	84.6
BERTje base	70.3	93.3	85.2
RobBERT-2026	69.6	93.3	82.4
RobBERT-2023 base	68.6	93.3	86.3

- Level on sentiment, mid-field on QA
- NLI is the **weak spot**: below every listed model
- Flat at ~ 82 for every token budget and data mix we tried, so it is not scale

A retriever on clean data: real passages, synthetic queries

RobBERT-2026 Retrieval, trained on cc-nl-retrieval



Passages are 92% Wikipedia, the rest EU law and tenders. Only the **queries** are synthetic.

One passage, many ways to search for it

Het seizoen 2018-19 van Feyenoord Basketbal [...] Het is het eerste seizoen als deel van de omnisportvereniging Feyenoord, hierdoor veranderde de naam van de club [...]

fact-check, chatty, short feyenoord basketbal nieuwe naam?

yes/no, formal, long Betreft het seizoen 2018-2019 de eerste jaargang waarin de Rotterdamse basketbalclub participeert onder de naam Feyenoord Basketbal, als onderdeel van de overkoepelende omnisportvereniging?

how-to, chatty, long hee ff snel, kan je me vertellen hoe het seizoen 2018-19 van feyenoord basketbal eruit zag, ben benieuwd naar de spelers en wie de coach was [...]

One passage, many ways to search for it

Het seizoen 2018-19 van Feyenoord Basketbal [...] Het is het eerste seizoen als deel van de omnisportvereniging Feyenoord, hierdoor veranderde de naam van de club [...]

fact-check, chatty, short feyenoord basketbal nieuwe naam?

yes/no, formal, long Betreft het seizoen 2018-2019 de eerste jaargang waarin de Rotterdamse basketbalclub participeert onder de naam Feyenoord Basketbal, als onderdeel van de overkoepelende omnisportvereniging?

how-to, chatty, long hee ff snel, kan je me vertellen hoe het seizoen 2018-19 van feyenoord basketbal eruit zag, ben benieuwd naar de spelers en wie de coach was [...]

- Type, register and length are sampled independently

One passage, many ways to search for it

Het seizoen 2018-19 van Feyenoord Basketbal [...] Het is het eerste seizoen als deel van de omnisportvereniging Feyenoord, hierdoor veranderde de naam van de club [...]

fact-check, chatty, short feyenoord basketbal nieuwe naam?

yes/no, formal, long Betreft het seizoen 2018-2019 de eerste jaargang waarin de Rotterdamse basketbalclub participeert onder de naam Feyenoord Basketbal, als onderdeel van de overkoepelende omnisportvereniging?

how-to, chatty, long hee ff snel, kan je me vertellen hoe het seizoen 2018-19 van feyenoord basketbal eruit zag, ben benieuwd naar de spelers en wie de coach was [...]

- Type, register and length are sampled independently
- Half are written as one of 5,000 Census-grounded [Belgian personas](#)

Retrieval: behind the big multilingual models

BEIR-NL, zero-shot nDCG@10

Model	NFCorpus	SciFact	ArguAna	SciDocs	NQ
gte-multilingual-base (305M)	27.97 [†]	64.41	52.85 [†]	15.86	47.42 [†]
multilingual-e5-base (278M)	24.17	67.23	47.06	10.53	36.06 [†]
RobBERT-2026 Retrieval (275M)	17.83	40.74	33.72	9.01	20.26
LaBSE (470M)	13.54	39.07	39.15	6.32	11.24
mContriever	13.36	37.89	39.60	4.93	10.50

Other models from Banar et al. (2025); [†] likely contaminated: probably trained on that dataset

- mE5 and gte are trained on far more data, often in-domain

Retrieval: behind the big multilingual models

BEIR-NL, zero-shot nDCG@10

Model	NFCorpus	SciFact	ArguAna	SciDocs	NQ
gte-multilingual-base (305M)	27.97 [†]	64.41	52.85 [†]	15.86	47.42 [†]
multilingual-e5-base (278M)	24.17	67.23	47.06	10.53	36.06 [†]
RobBERT-2026 Retrieval (275M)	17.83	40.74	33.72	9.01	20.26
LaBSE (470M)	13.54	39.07	39.15	6.32	11.24
mContriever	13.36	37.89	39.60	4.93	10.50

Other models from Banar et al. (2025); [†] likely contaminated: probably trained on that dataset

- mE5 and gte are trained on far more data, often in-domain
- Ahead of LaBSE and mContriever on 4 of 5 tasks; behind on ArguAna

Lessons

What I would tell myself in 2019

1. The `tokenizer` is what makes a model Dutch, and you can swap it without starting over

What I would tell myself in 2019

1. The **tokenizer** is what makes a model Dutch, and you can swap it without starting over
2. Language moves: plan for updates instead of patching BPE vocabularies

What I would tell myself in 2019

1. The **tokenizer** is what makes a model Dutch, and you can swap it without starting over
2. Language moves: plan for updates instead of patching BPE vocabularies
3. Strong multilingual bases shrink the gain of continued pretraining

What I would tell myself in 2019

1. The **tokenizer** is what makes a model Dutch, and you can swap it without starting over
2. Language moves: plan for updates instead of patching BPE vocabularies
3. Strong multilingual bases shrink the gain of continued pretraining
4. Data you **control** beats data you scrape: seeds, licences, synthetic amplification

Thank you!

Models and data

- `pdelobelle/robbert-2026` and `robbert-2026-retrieval`
- `PleIAs/SYNTH` and `PleIAs/Baguettotron`
- `DTAI-KULeuven/robbert-2023-dutch-large`

Read more

- `pieter.ai/robbert`
- `arxiv.org/abs/2609.37891` (Synth)

