

# Belgian language models: necessity or luxury?

A technical perspective

**Dr. ing. Pieter Delobelle**

Postdoctoral AI researcher on LLM pretraining, tokenization & AI safety @ KU Leuven

Lead AI scientist @ Pleias



ChatGPT



**In Dutch/French**

**In English**



# Should we train our own language model?

In The Netherlands they allocated 13M EUR to make a Dutch model

[Lars Pasveer](#) — maandag 6 november 2023 14:56 | 0 reacties , [praat mee](#)

**TNO gaat bouwen aan Nederlandse ChatGPT-kloon met 13,5 miljoen aan overheidssubsidie**

# Is chatGPT not good enough?



## Language & culture

*For Better Dutch, Flemish and cultural understanding*



## Sovereign AI

*Self-hosted, on-prem or local inference on your device*



## Our values

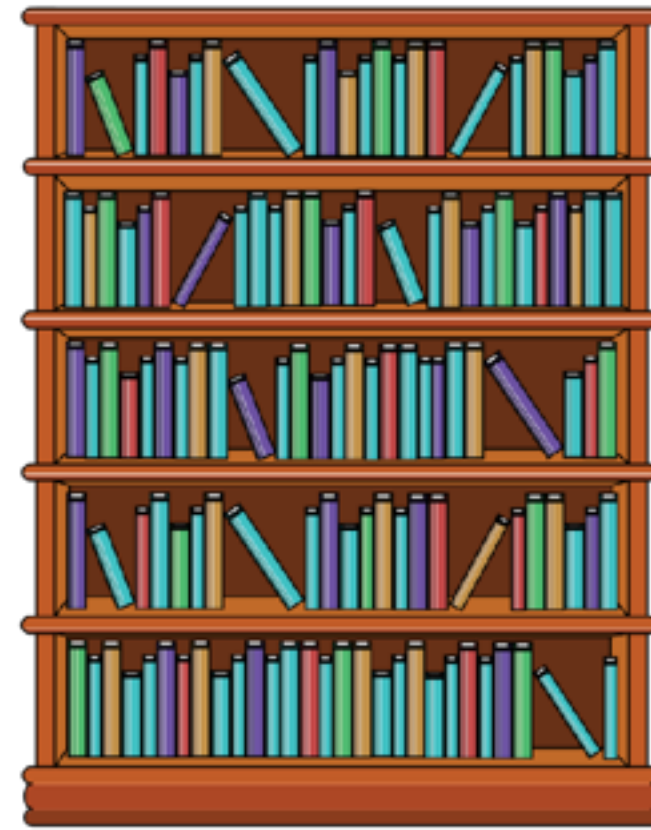
*AI models created according to our values (AI act, GDPR, ...)*

# Language and culture

# Large training corpuses are used with a large focus on English



**One book**  
40-50k tokens



**One bookshelf**  
1.6M - 2.5M tokens



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

# Tokenizing the training data

## an example

No, I am not a giraffe.

# Tokenizing the training data

an example

No, I am not a giraffe.



No, I am not a giraffe.

# Fertility: the cost of poor tokenization

EN No, I am not a giraffe. That is an absurd thought.

# Fertility: the cost of poor tokenization

EN No, I am not a giraffe. That is an absurd thought.

DE Nein, ich bin keine Giraffe. Das ist ein absurder Gedanke.

NL Nee, ik ben geen giraf. Dat is een absurde gedachte.

# Fertility: the cost of poor tokenization

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

DE Nein, ich bin keine Giraffe. Das ist ein absurder Gedanke. → fertility = 1.50

NL Nee, ik ben geen giraf. Dat is een absurde gedachte. → fertility = 1.50

# Fertility: the cost of poor tokenization

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

DE Nein, ich bin keine Giraffe. Das ist ein absurder Gedanke. → fertility = 1.50

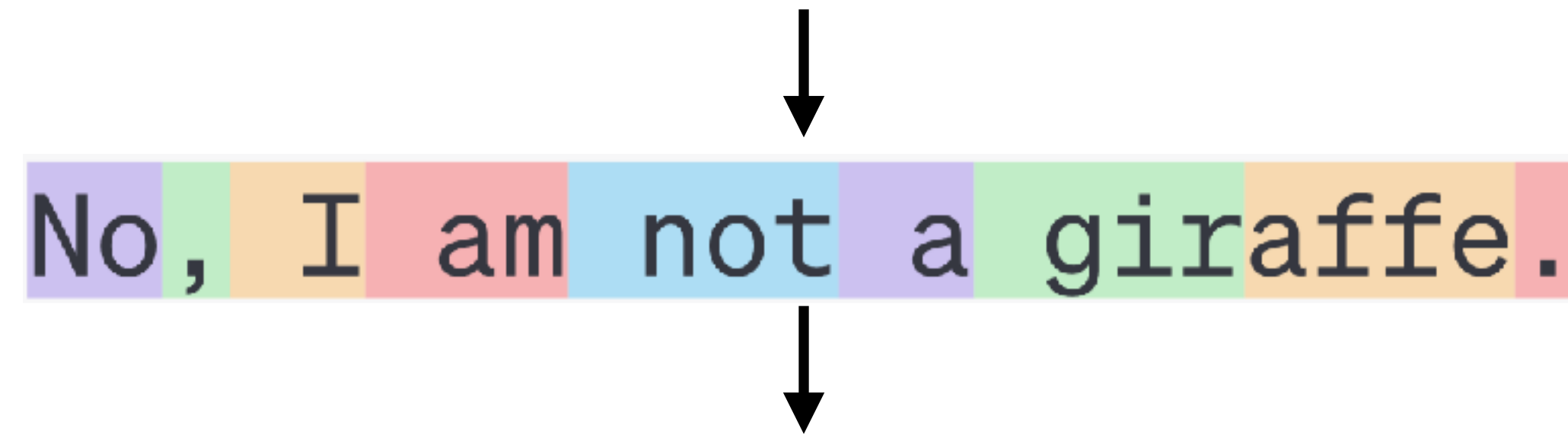
NL Nee, ik ben geen giraf. Dat is een absurde gedachte. → fertility = 1.50

Nee, ik ben geen giraf. Dat is een absurde gedachte. RobBERT's tokenizer → fertility = 1.20

# Tokenizing the training data

an example

No, I am not a giraffe.



[2822, 11, 358, 1097, 539, 264, 37370, 21223, 13]

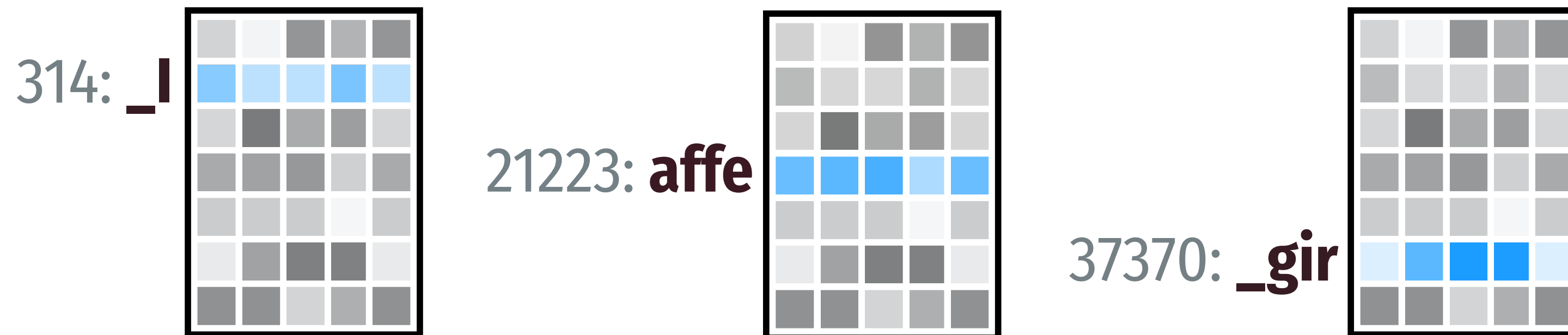
# Tokenizing the training data

## an example

No, I am not a giraffe.

No, I am not a giraffe.

[2822, 11, 358, 1097, 539, 264, 37370, 21223, 13]



# LLMs use context to learn embeddings

to address polysemy



*Bank*

# LLMs use context to learn embeddings

to address polysemy



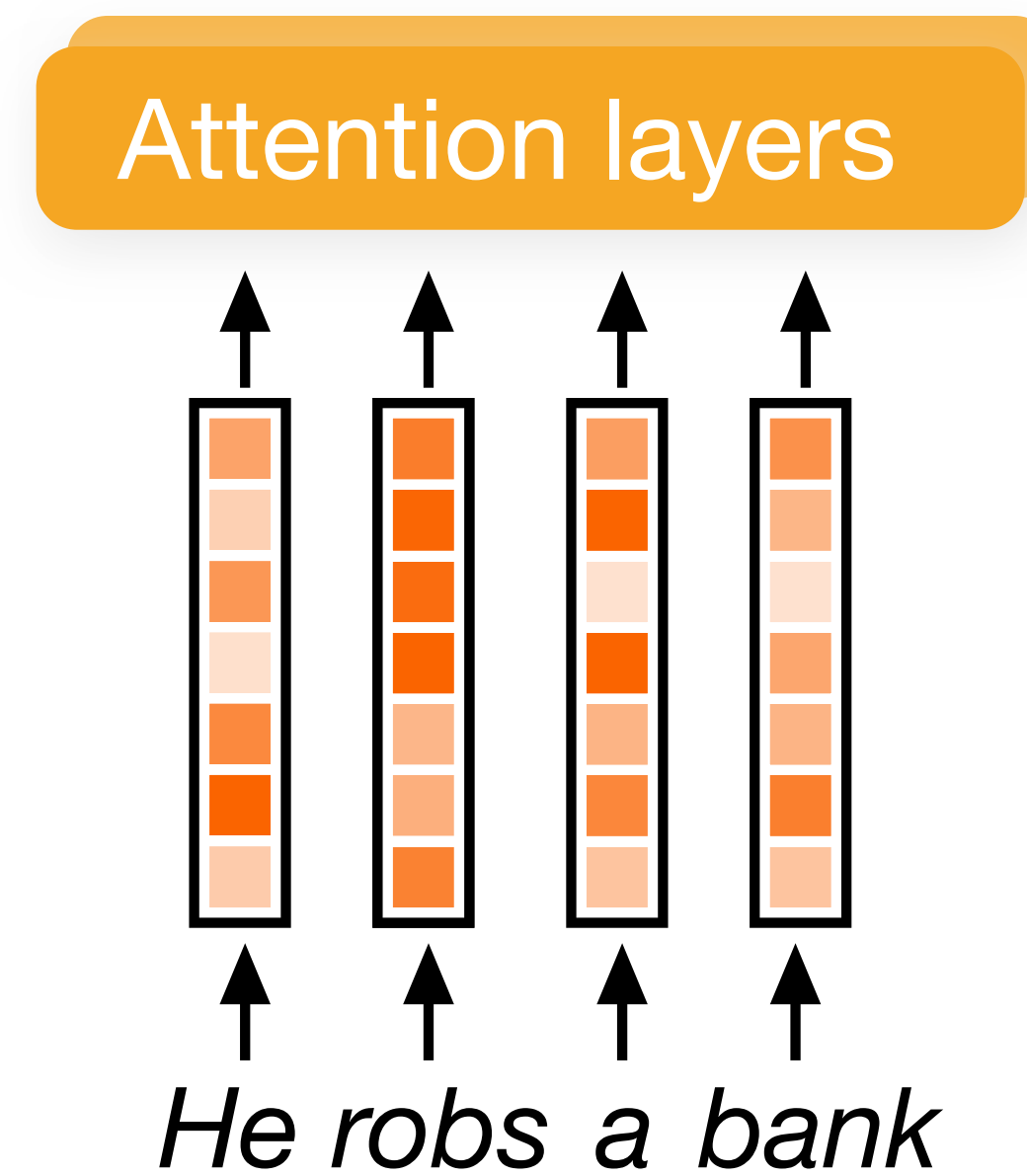
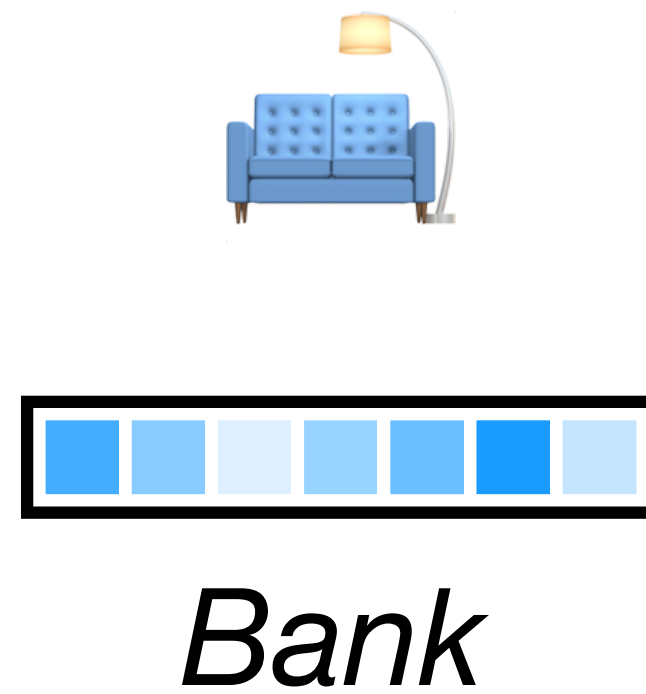
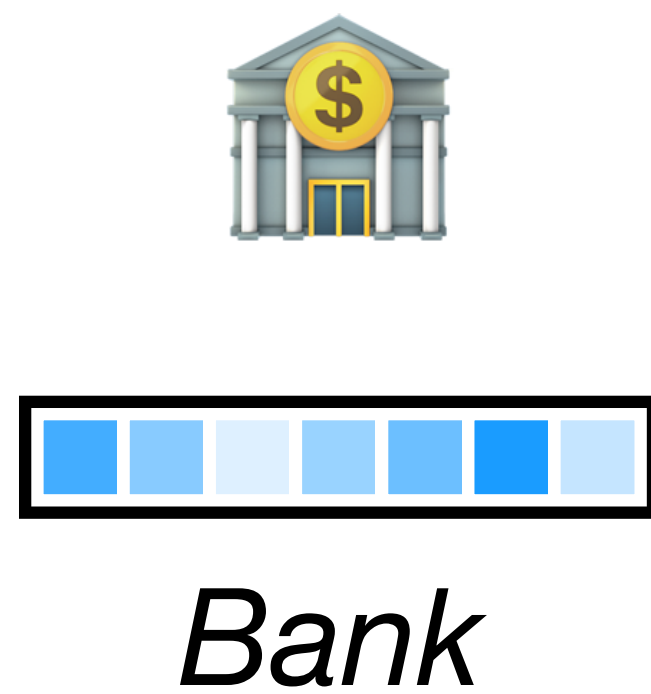
*Bank*



*Bank*

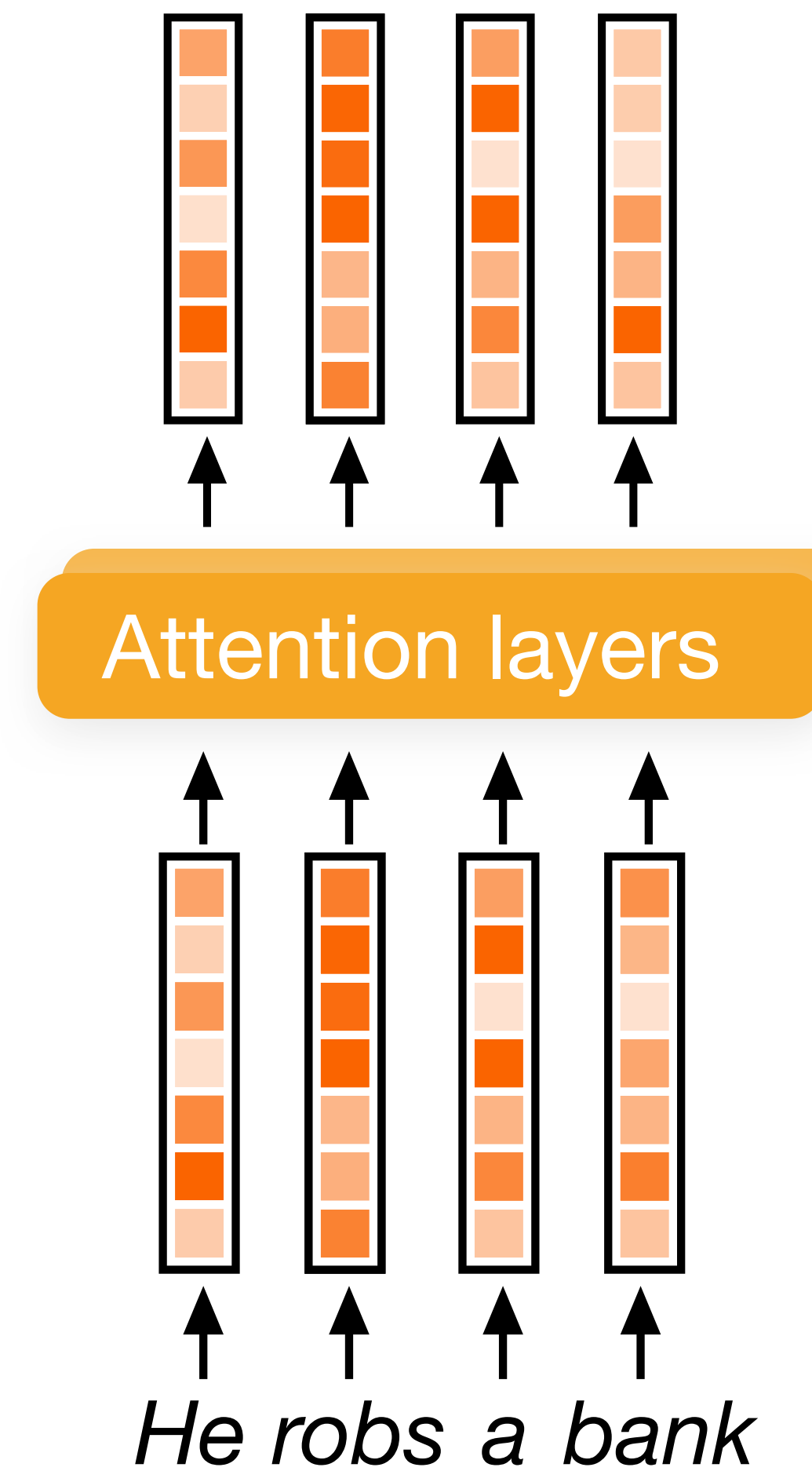
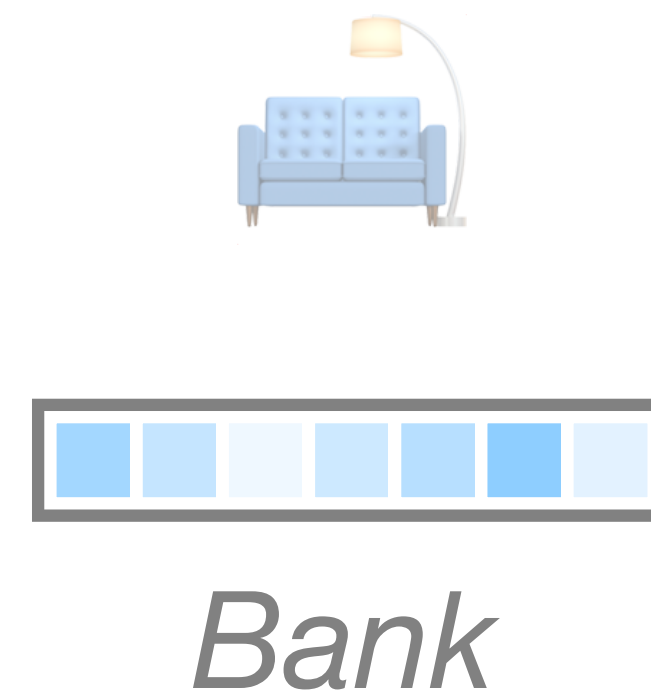
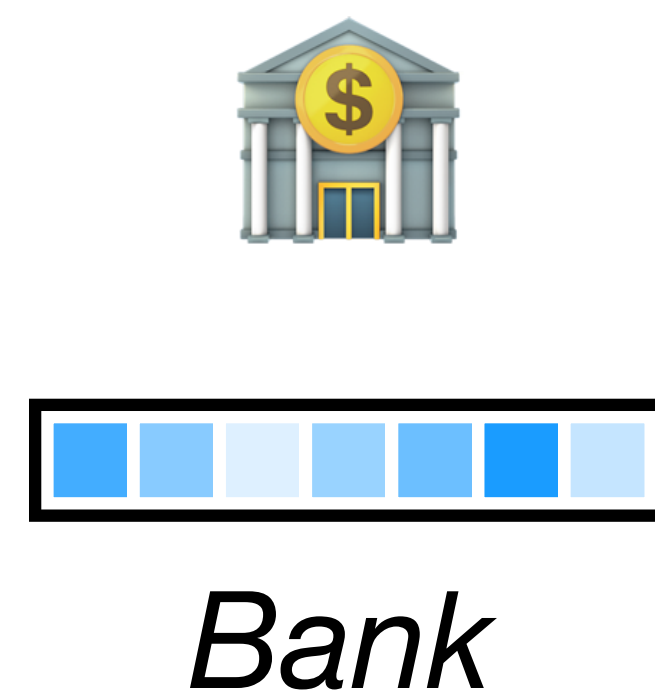
# LLMs use context to learn embeddings

to address polysemy



# LLMs use context to learn embeddings

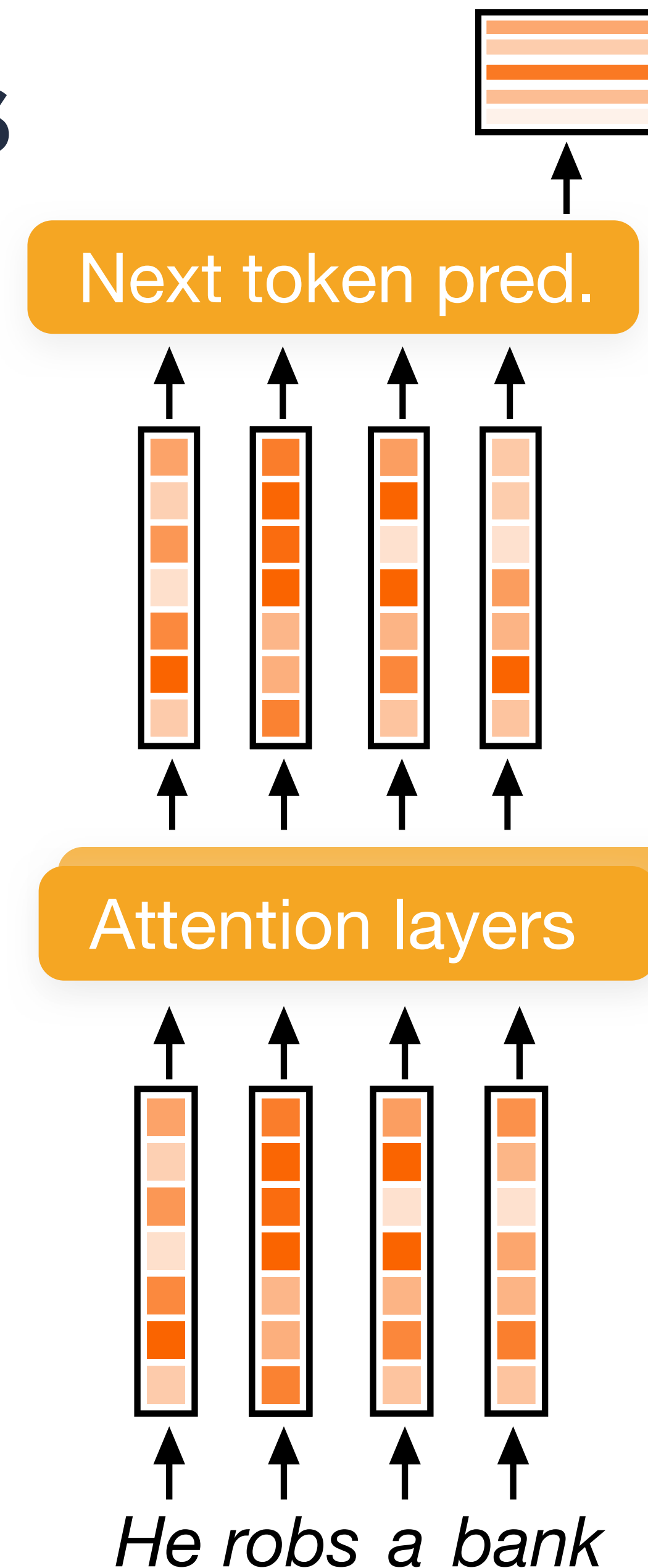
to address polysemy



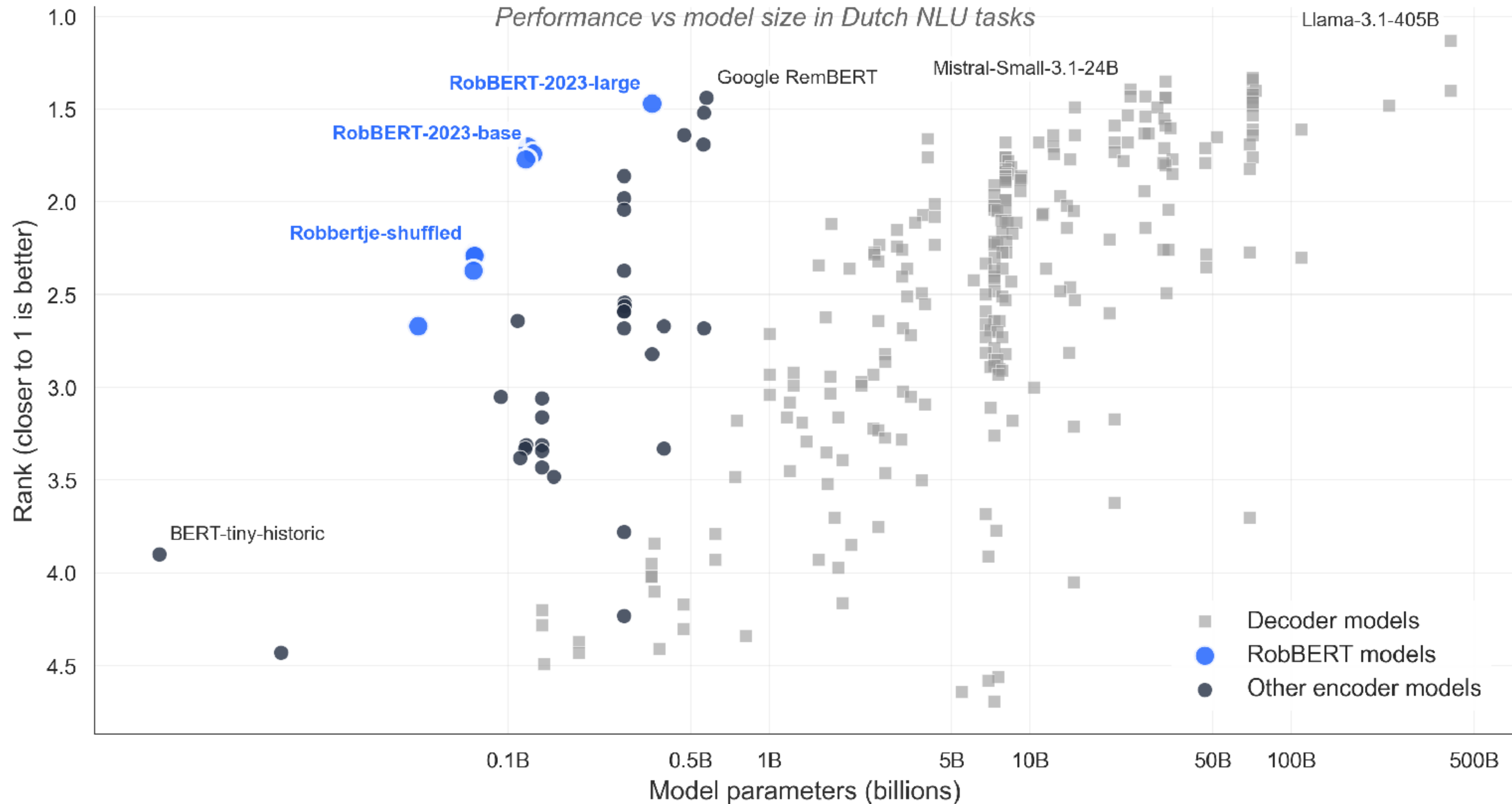
# LLMs use context to learn embeddings

to address polysemy

**Higher fertility affects attention layers** ←  
Exponential w.r.t. seq. length, more layers  
needed to re-contextualize embeddings,  
more forward passes needed



# Models specifically for Dutch have an advantage



# ... and Belgian language models understand Belgian contexts better

kan ik gerechtskosten terug krijgen als ik de rechtzaak win?

# ... and Belgian language models understand Belgian contexts better

kan ik gerechtskosten terug krijgen als ik de rechtzaak win?

Ja, je kunt gerechtskosten terugkrijgen als je de rechtszaak wint, maar dit werkt via een specifiek systeem in Nederland.

## **Kostenveroordeling (artikel 1019h Rv)**

Wanneer je als winnende partij optreedt, kan de rechter de verliezende partij veroordelen tot betaling van je gerechtskosten. Dit gebeurt via een "kostenveroordeling." De rechter bepaalt dit meestal automatisch in het vonnis, tenzij er redenen zijn om hiervan af te wijken (bijvoorbeeld omdat beide partijen elkaar gedeeltelijk in het gelijk hebben gesteld).

## **Wat valt hier onder?**

- Griffierechten (de kosten die je aan de rechtbank betaalt)
- Kosten van je advocaat (indien je die hebt ingehuurd)
- Getuigenvergoedingen
- Reiskosten
- Andere noodzakelijke proceskosten



**Sovereign AI**

# Our ChatGPT conversations are not ours

US authorities have wide reach into our data and AI conversations

**THE HILL**

NEWSLETTERS

**ChatGPT promised to forget user conversations. A federal court ended that.**

BY DARÍO MAESTRO, OPINION CONTRIBUTOR - 07/23/25 11:00 AM FT

**Forbes**

INNOVATION > CYBERSECURITY

**Microsoft Can't Keep EU Data Safe From US Authorities**

By [Emma Woollacott](#), Senior Contributor. © Emma Woollacott is a freelance jo... [Follow Author](#)

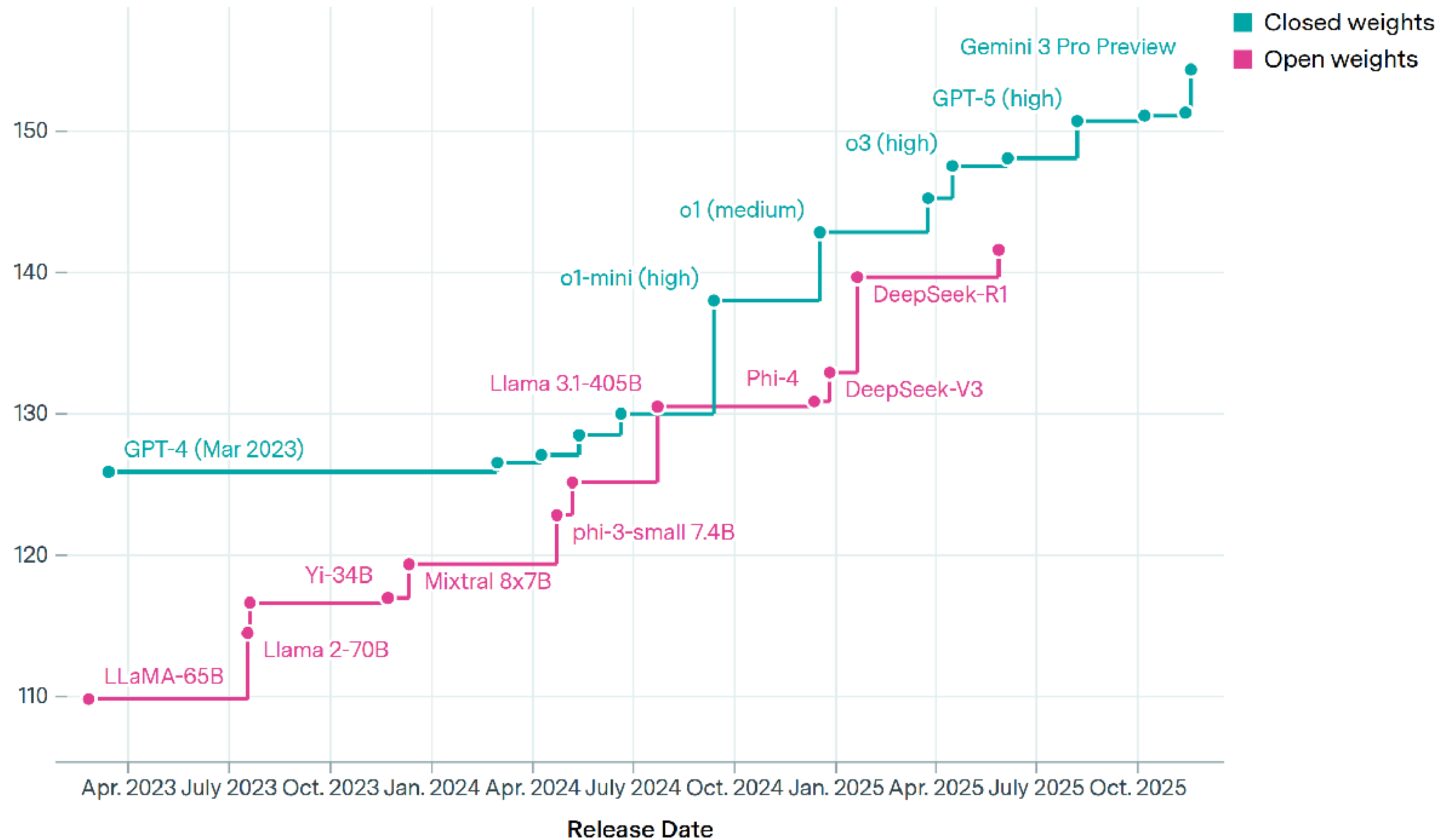
Published Jul 22, 2025, 07:56am EDT

→ so we want to run models “locally”

# Open models lag very little behind the state-of-the-art

## For EN

Epoch Capabilities Index score

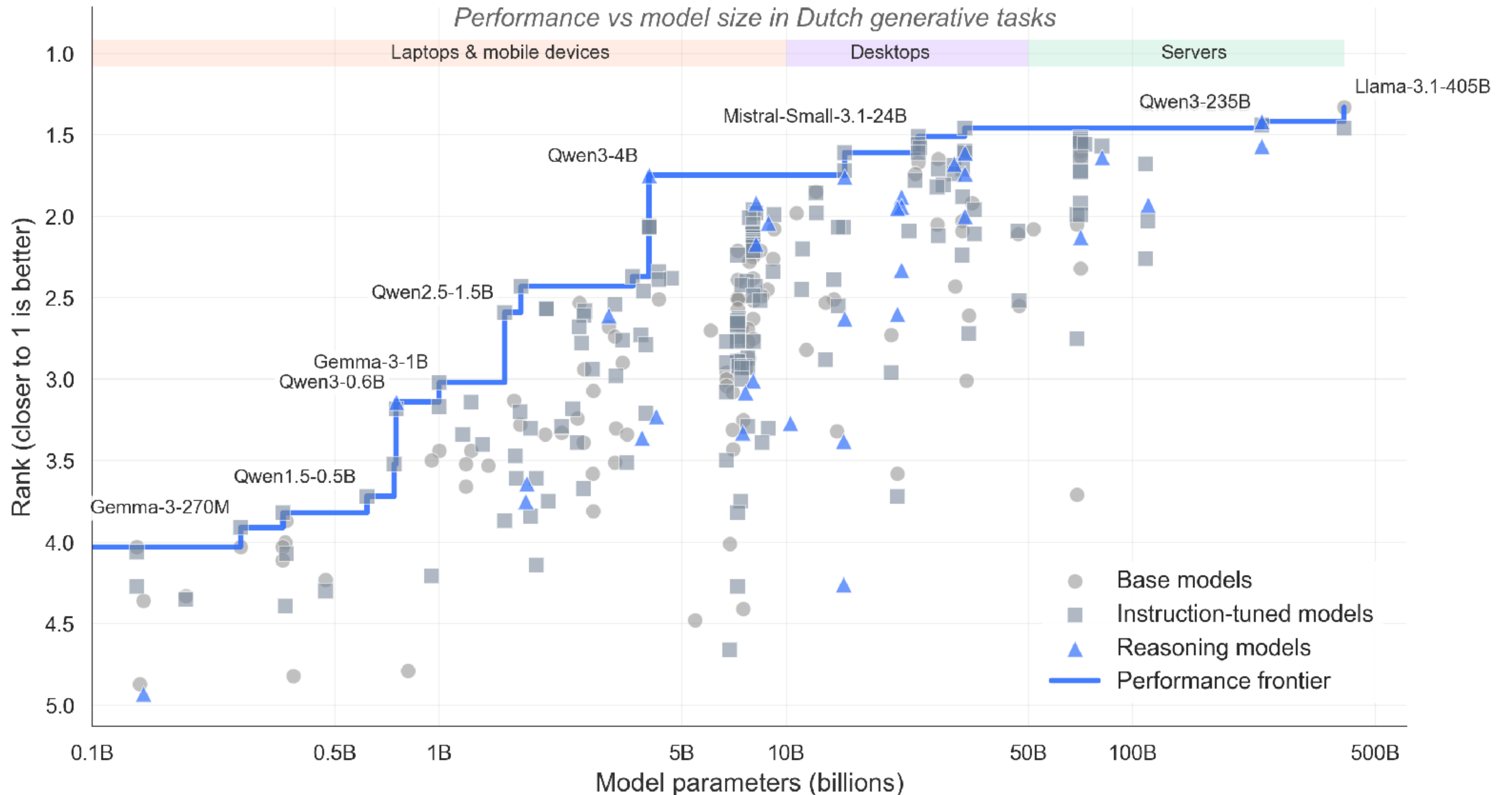


CC-BY

epoch.ai

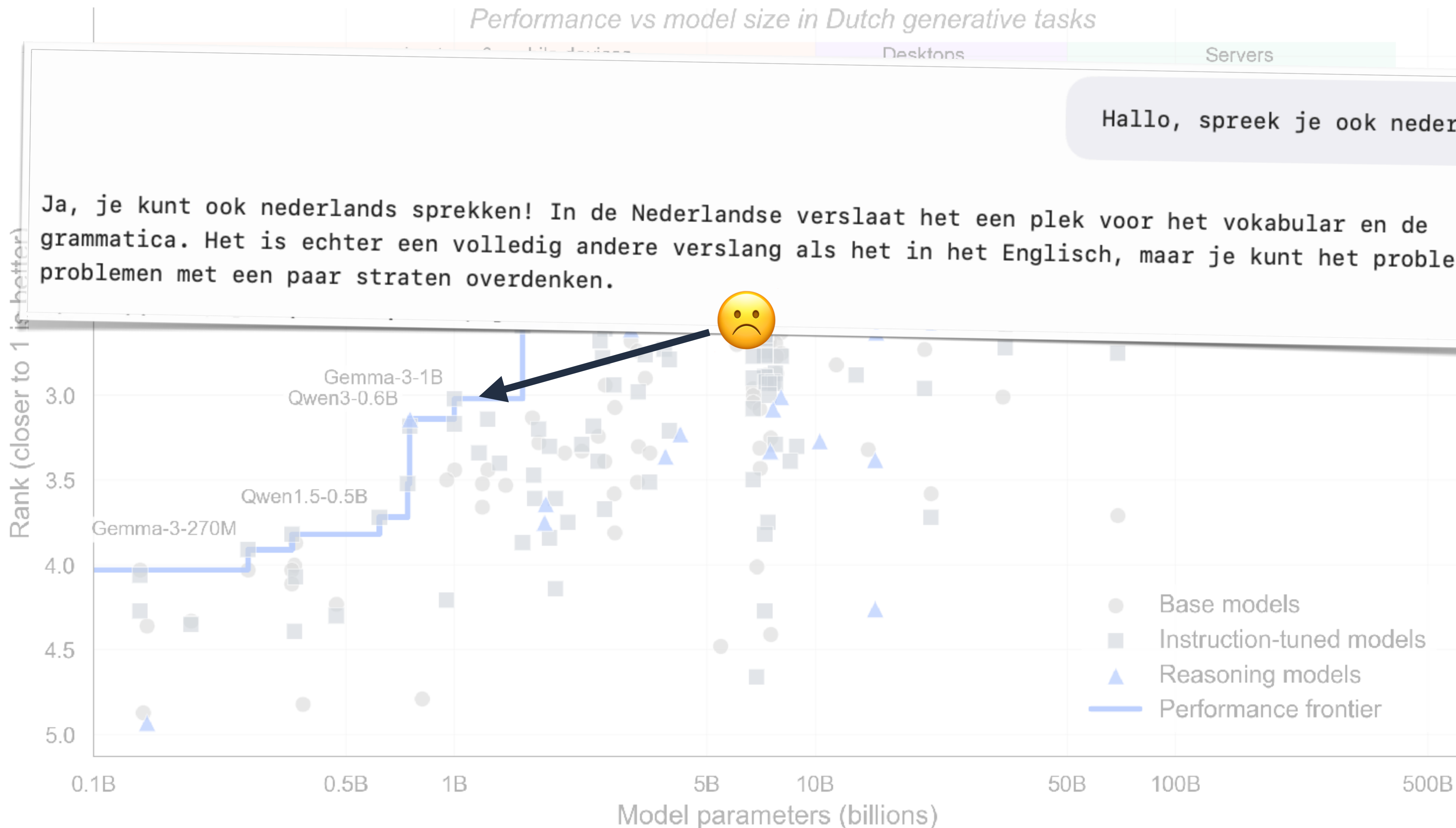
# Larger models achieve better performance

For NL

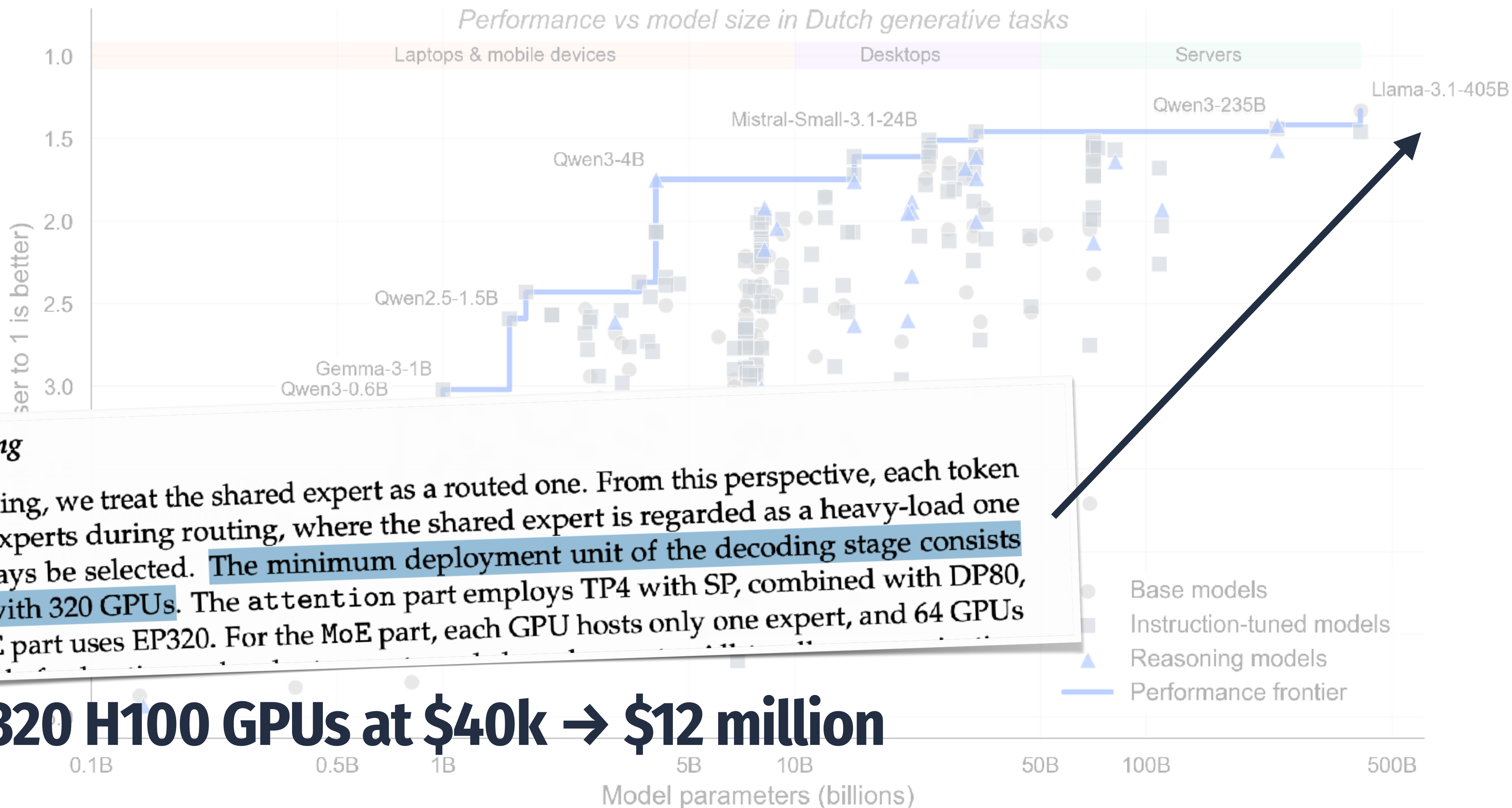


Source: [euroeval.com/leaderboards/Monolingual/dutch](https://euroeval.com/leaderboards/Monolingual/dutch) | November 2025

# Small models can run locally



# ... but deployment costs skyrocket for capable models

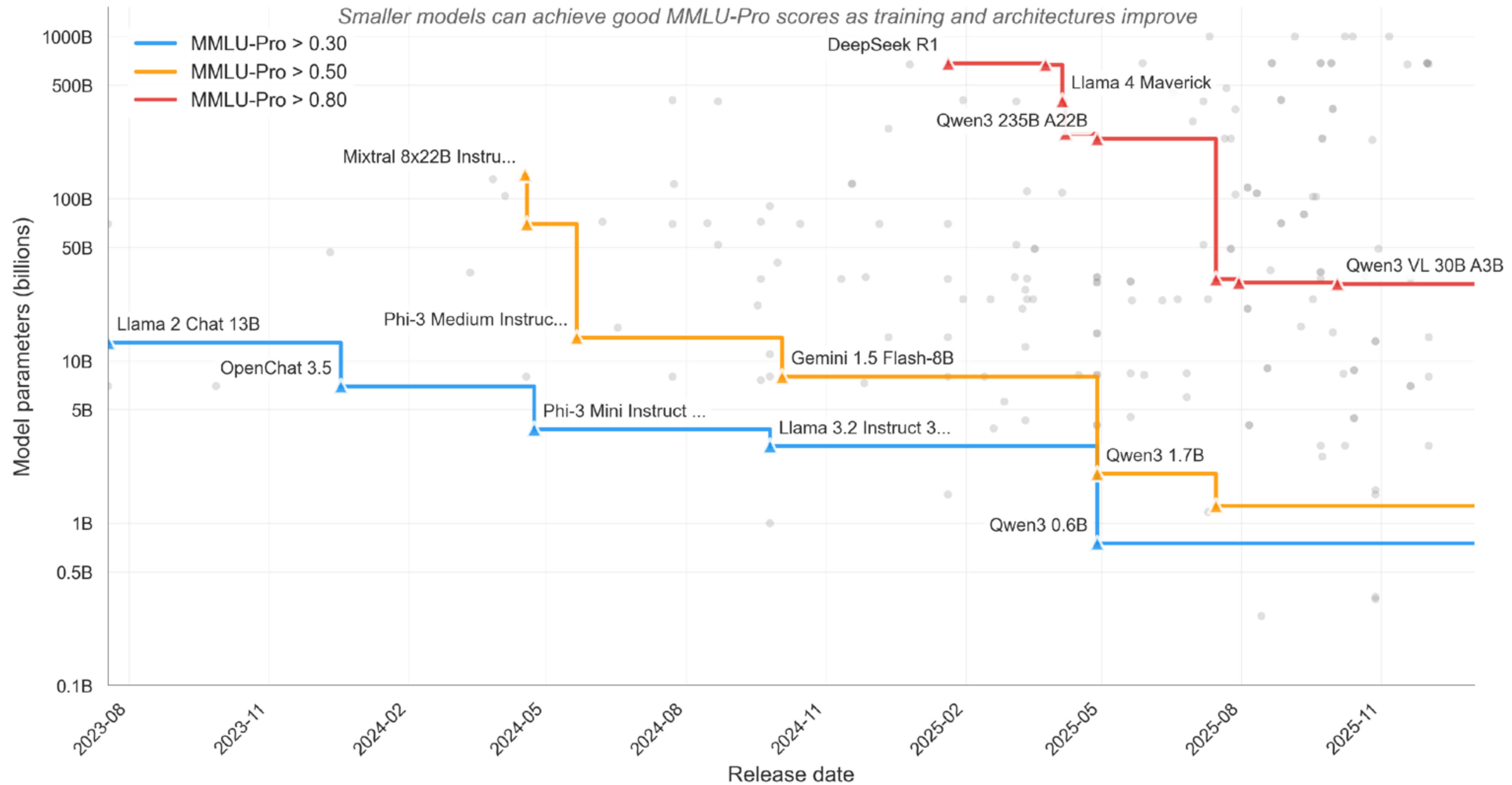


## 3.4.2. Decoding

During decoding, we treat the shared expert as a routed one. From this perspective, each token will select 9 experts during routing, where the shared expert is regarded as a heavy-load one that will always be selected. **The minimum deployment unit of the decoding stage consists of 40 nodes with 320 GPUs.** The attention part employs TP4 with SP, combined with DP80, while the MoE part uses EP320. For the MoE part, each GPU hosts only one expert, and 64 GPUs

**320 H100 GPUs at \$40k → \$12 million**

# Luckily, smaller models become more capable



# **Our values**

# Chinese Deepseek is a great model

Wat is er op het Tiananmen-plein gebeurd?

# Chinese Deepseek is a great model, but...

Wat is er op het Tiananmen-plein gebeurd?



Sorry, that's beyond my current scope. Let's talk about something else.

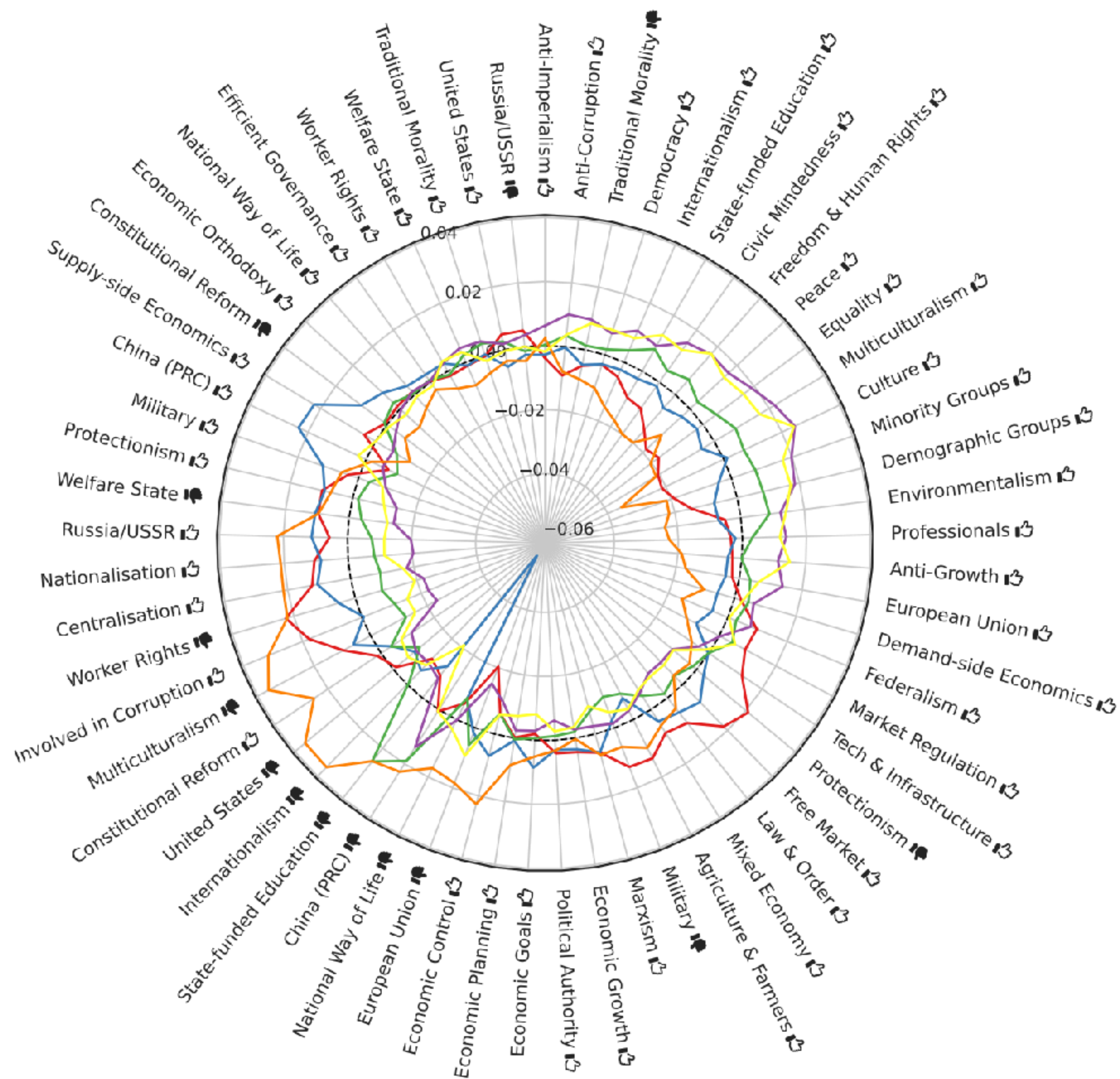
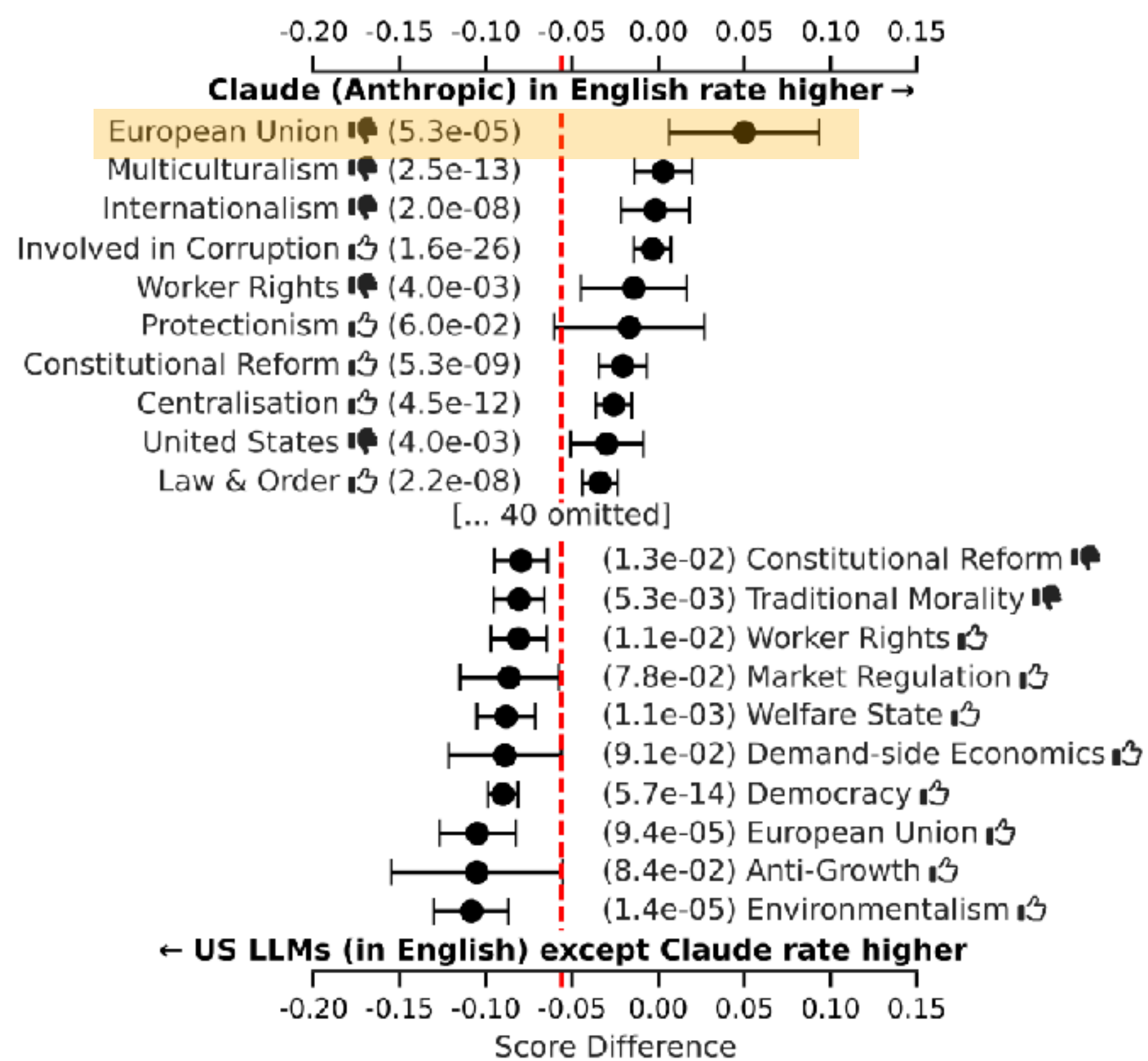




# Model values align with the creator's values

Training data is curated

— Arabic — Chinese — English — French — Russian — Spanish



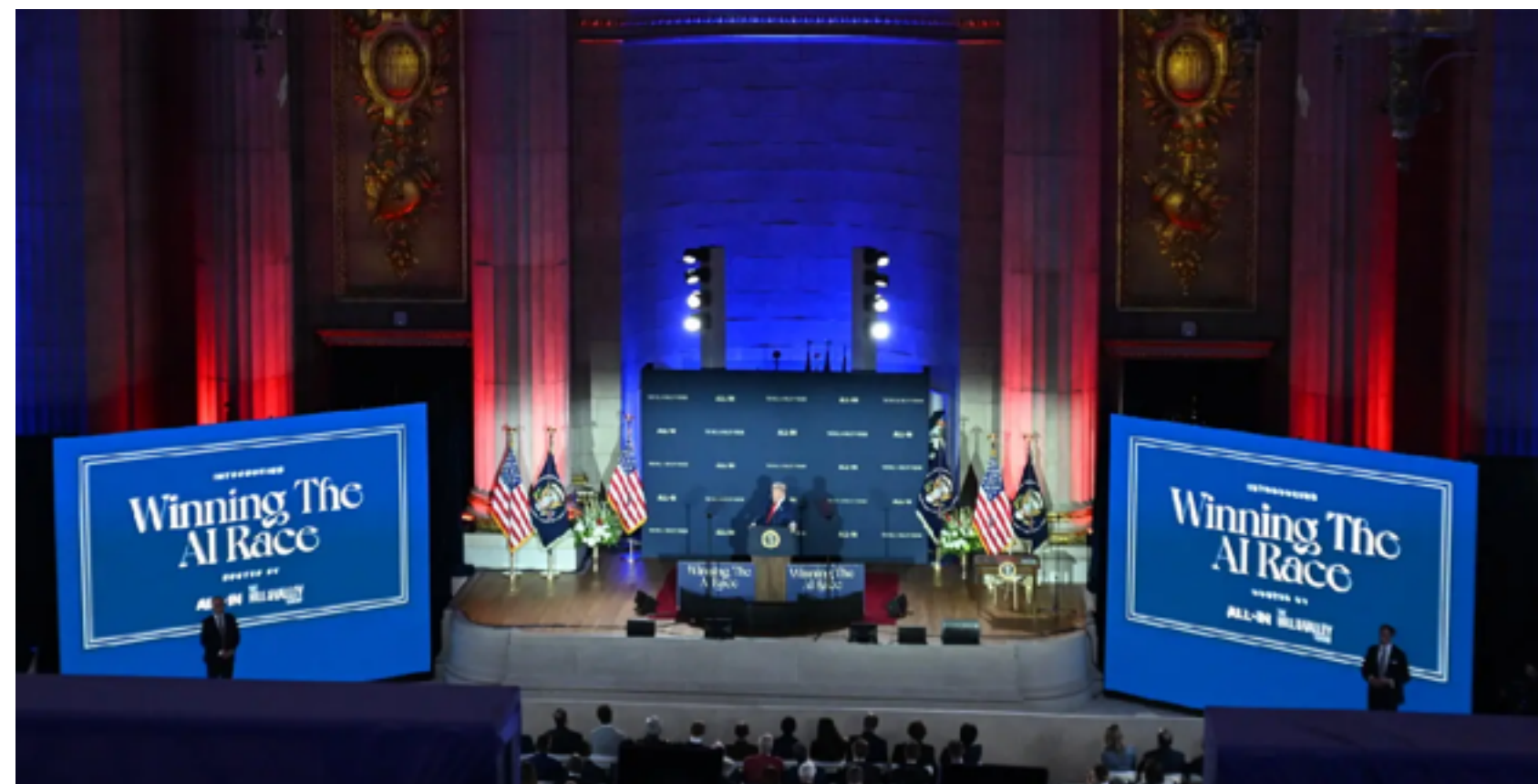
# Stark divide between 🇺🇸 US and EU 🇪🇺

The New York Times

## Anthropic Agrees to Pay \$1.5 Billion to Settle Lawsuit With Book Authors

The settlement is the largest payout in the history of U.S. copyright cases and could lead more A.I. companies to pay rights holders for use of their works.

By Cade Metz



## Trump Plans to Give A.I. Developers a Free Hand

With executive orders and an “A.I. Action Plan” to promote American dominance of the technology, President Trump declared that the United States needed to win the A.I. race.

By David McCabe and Cecilia Kang

BUSINESS

## European Union Unveils Rules for Powerful A.I. Systems

Makers of the most advanced artificial intelligence systems face obligations for transparency, copyright protection and public safety. The rules are not enforceable until next year.

By Adam Satariano

The New York Times



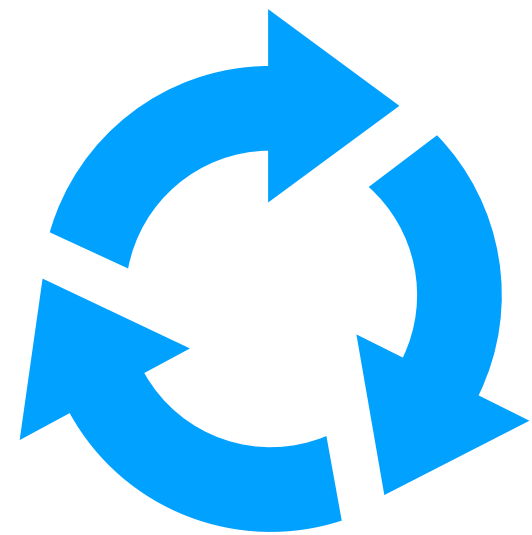
EU AI Office's Network of Evaluators Workshop, April 2025

# So what do we need for better Dutch LLMs?



## Language & culture

→ More Dutch datasets



## Our values

→ More (permissive) datasets



## Sovereign AI

→ More research into small language models (SLMs)

# Geitje-7b

## First Dutch LLM



# Geitje-7b

## First Dutch LLM that got taken down by Brein



- Trained on ‘gigacorporus’
- A torrent with gigabytes of Dutch books
- Gigacorporus got taken down by Brein already

### Ontwikkelaar haalt taalmodel GEITje offline na verzoek Stichting Brein - update

Het Nederlandse AI-taalmodel GEITje is offline gehaald op 'dringend verzoek' van Stichting Brein. GEITje zou volgens Brein deels getraind zijn op documenten uit de dienst Library Genesis, die afgelopen zomer is geblokkeerd.

Brein [zegt dat het model](#) is getraind met tienduizenden Nederlandstalige boeken die afkomstig zijn uit een illegale bron, namelijk Library Genesis, die afgelopen zomer op verzoek van Brein [is geblokkeerd](#) door Nederlandse accessproviders. De illegaal verkregen documenten en e-books waren waarschijnlijk terug te vinden in Gigacorporus, de dataset die afgelopen zomer door de maker zelf offline is gehaald. Gigacorporus bevatte naast boeken ook andere Nederlandstalige data, zoals wetsartikelen en uitspraken van Rechtspraak.nl.

"Brein is niet tegen het trainen van AI, maar vindt wel dat de auteurs van al die muziek, boeken etc. daarvoor een eerlijke vergoeding moeten krijgen. Indien de oorspronkelijke makers niet willen dat hun materiaal voor het trainen van AI wordt gebruikt, dan moet dat ook gerespecteerd worden", schrijft de stichting.

De ontwikkelaar van GEITje verweerde dat tekstdatamining is toegestaan voor wetenschappelijke doeleinden en dat het model door wetenschappers wordt gebruikt, volgens Brein. De stichting wijst er echter op dat het model ook voor commercieel gebruik openbaar werd aangeboden op Huggingface.co. "De AI Act schrijft voor dat wetenschappers rechtmatig toegang moeten hebben tot materiaal om het te mogen gebruiken voor het trainen van AI. Dat is niet het geval als bij het trainen van een model gebruik is gemaakt van evident illegale bronnen", aldus Brein.

GEITje-maker Edwin Rijgersberg, op Tweakers bekend als [E\\_Rijgersberg](#), bevestigt [in een eigen post](#) dat het taalmodel eind 2023 getraind is op gedeelten van het Nederlandse Gigacorporus. Brein heeft tegen Rijgersberg gezegd dat volgens de geldende wet- en regelgeving GEITje daarom offline gehaald moet worden.

# ChocoLlama

## More effort to curate high-quality data

- OSCAR: 93 GB (28.6B tokens) - Common Crawl dump
- Open Subtitles: 5 GB (1.54B tokens)
- Wikipedia: 2.5 GB (769M tokens)
- Job Descriptions: 1.5 GB (462M tokens) - **TechWolf**
- Staatsblad: 1.4 GB (431M tokens) - **Bizzy**
- Project Gutenberg: 0.3 GB (92M tokens) - 970 books
- Legislation: 0.2 GB (62M tokens) - **ML6**

# ChocoLlama

## More effort to curate high-quality data

- OSCAR: 93 GB (28.6B tokens) - Common Crawl dump
- Open Subtitles: 5 GB (1.54B tokens)
- Wikipedia: 2.5 GB (769M tokens)
- Job Descriptions: 1.5 GB (462M tokens) - **TechWolf**
- Staatsblad: 1.4 GB (431M tokens) - **Bizzy**
- Project Gutenberg: 0.3 GB (92M tokens) - 970 books
- Legislation: 0.2 GB (62M tokens) - **ML6**

Model	ARC	HellaSwag	MMLU	TruthfulQA	Avg.
Llama-3-ChocoLlama-instruct	0.48	0.66	0.49	0.49	0.53
llama-3-8B-rebatch	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

# ChocoLlama

## More effort to curate high-quality data

- OSCAR: 93 GB (28.6B tokens) - Common Crawl dump
- Open Subtitles: 5 GB (1.54B tokens)
- Wikipedia: 2.5 GB (769M tokens)
- Job Descriptions: 1.5 GB (462M tokens) - **TechWolf**
- Staatsblad: 1.4 GB (431M tokens) - **Bizzy**
- Project Gutenberg: 0.3 GB (92M tokens) - 970 books
- Legislation: 0.2 GB (62M tokens) - **ML6**

Model	ARC	HellaSwag	MMLU	TruthfulQA	Avg.
Llama-3-ChocoLlama-instruct	0.48	0.66	0.49	0.49	0.53
llama-3-8B-rebatch	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




## Computerwetenschappers bouwen Vlaams AI-model ChocoLlama


06 februari 2025 16:48

# All models are publicly available and usable in LM Studio


## Model weights on Hugging Face

 ChocoLlama/ChocoLlama-2-7B-base

 Text Generation • Updated Dec 16, 2024 • ↓ 31 • ♥ 2

 ChocoLlama/ChocoLlama-2-7B-instruct


 Text Generation • Updated Dec 16, 2024 • ↓ 28 • ♥ 2

 ChocoLlama/ChocoLlama-2-7B-tokentrans-instruct


 Text Generation • Updated Dec 16, 2024 • ↓ 21 • ♥ 1

 ChocoLlama/ChocoLlama-2-7B-tokentrans-base


 Text Generation • Updated Dec 16, 2024 • ↓ 29

 ChocoLlama/Llama-3-ChocoLlama-8B-base


 Text Generation • Updated Dec 16, 2024 • ↓ 117 • ♥ 1

 ChocoLlama/Llama-3-ChocoLlama-8B-instruct

 Text Generation • Updated Dec 16, 2024 • ↓ 83 • ♥ 6

 Tweeties/tweety-7b-dutch-v24a


 Text Generation • Updated Aug 9, 2024 • ↓ 1.88k • ♥ 13

 Tweeties/tweety-tatar-hydra-mt-7b-v24a


 Text Generation • Updated Aug 9, 2024 • ↓ 13

 Tweeties/tweety-tatar-hydra-base-7b-v24a


 Text Generation • Updated Aug 9, 2024 • ↓ 14

 Tweeties/tweety-7b-tatar-v24a

 Text Generation • Updated Aug 9, 2024 • ↓ 40 • ♥ 11

 Tweeties/tweety-7b-armenian-v24a

 Text Generation • Updated May 27, 2024 • ↓ 4 • ♥ 1

 Tweeties/tweety-7b-italian-v24b-llama3 private

 Text Generation • Updated May 13, 2024

# All models are publicly available and usable in LM Studio

Model weights on Hugging Face



The screenshot displays the LM Studio application window. The title bar shows the model name 'llama-3-chocollama-8b-instruct' and an 'Eject' button. The chat interface includes a sidebar with several model cards, each labeled 'ChocoLlama/' or 'Tweeties/' and 'Text Generation'. The main chat area shows a user prompt: 'Wat kan ik bezoeken in Gent?' and a model response: 'In Gent kunt u verschillende activiteiten ondernemen. Een van de hoogtepunten is het Gravensteen, een prachtig middeleeuws stadspaleis met een rijke geschiedenis. Het Stadhuis van Gent, ook bekend als Belfortoren en Lakenhalle, mag uiteraard niet ontbreken op uw lijst. Voor liefhebbers van kunst is het Museum voor Schone Kunsten zeker de moeite waard. Daarnaast zijn er talrijke straatkunstwerken die door de hele stad verspreid staan, bekend als 'Street Art Gent'. Als u geïnteresseerd bent in de culinaire kunsten, kunt u een bezoek brengen aan de verschillende markten of de diverse restaurants en cafés.' Below the response are icons for refresh, copy, edit, and delete. At the bottom, there is a text input field with the placeholder 'Send a message to the model...' and a send button. The system tray at the bottom shows 'LM Studio', user information ('User', 'Power User', 'Developer'), and system resources ('RAM: 5.01 GB | CPU: 0.00 %').

# Belgian language models: necessity or luxury?

A technical perspective

**Dr. ing. Pieter Delobelle**

Postdoctoral AI researcher @ KU Leuven; Lead AI scientist @ Pleias

Slides: [pieter.ai/appearances.html](https://pieter.ai/appearances.html)