

AI is (nog) geen stagiair

Wat taalmodellen wel en niet kunnen

Dr. ing. Pieter Delobelle

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

Lead AI scientist @ Pleias



ChatGPT



**Who trusts AI for
important work?**

Who doesn't?



How large language models work

Data security & sovereignty

Model quality

Hallucinations and retrieval augmented generation

Our values

Closing thoughts & demo

Who am I?

- 2025-...** Postdoctoral researcher at KU Leuven
- 2026-...** Lead AI scientist at Pleias 🇫🇷
- 2024-2025** LLM engineer at Aleph Alpha (acc. by Cohere) 🇩🇪
- 2023** Apple 🇬🇧

Postdoc and PhD @ KU Leuven's DTAI research group

First author of our RobBERT model

state-of-the-art Dutch BERT language model, 4M+ downloads

Researching large language models training

Everything from tokenization, synthetic data, model training to LLM inference

Expert advisor for the EU's AI Act Code of Practice

and prev. inaugural member of the KU Leuven GenAI board

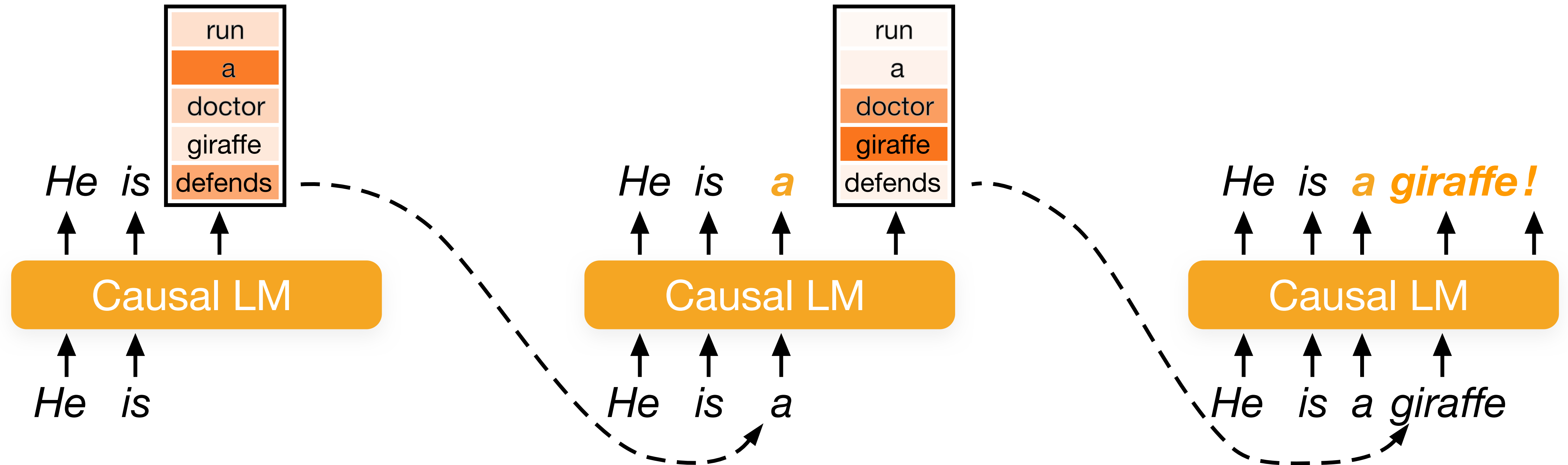
and technical advisor in a strategic litigation case against companion AIs



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

How large language models work

Generating text with LMs



Parts of a language models

'Heads' of a language model

How a model predicts the next word

Attention mechanism

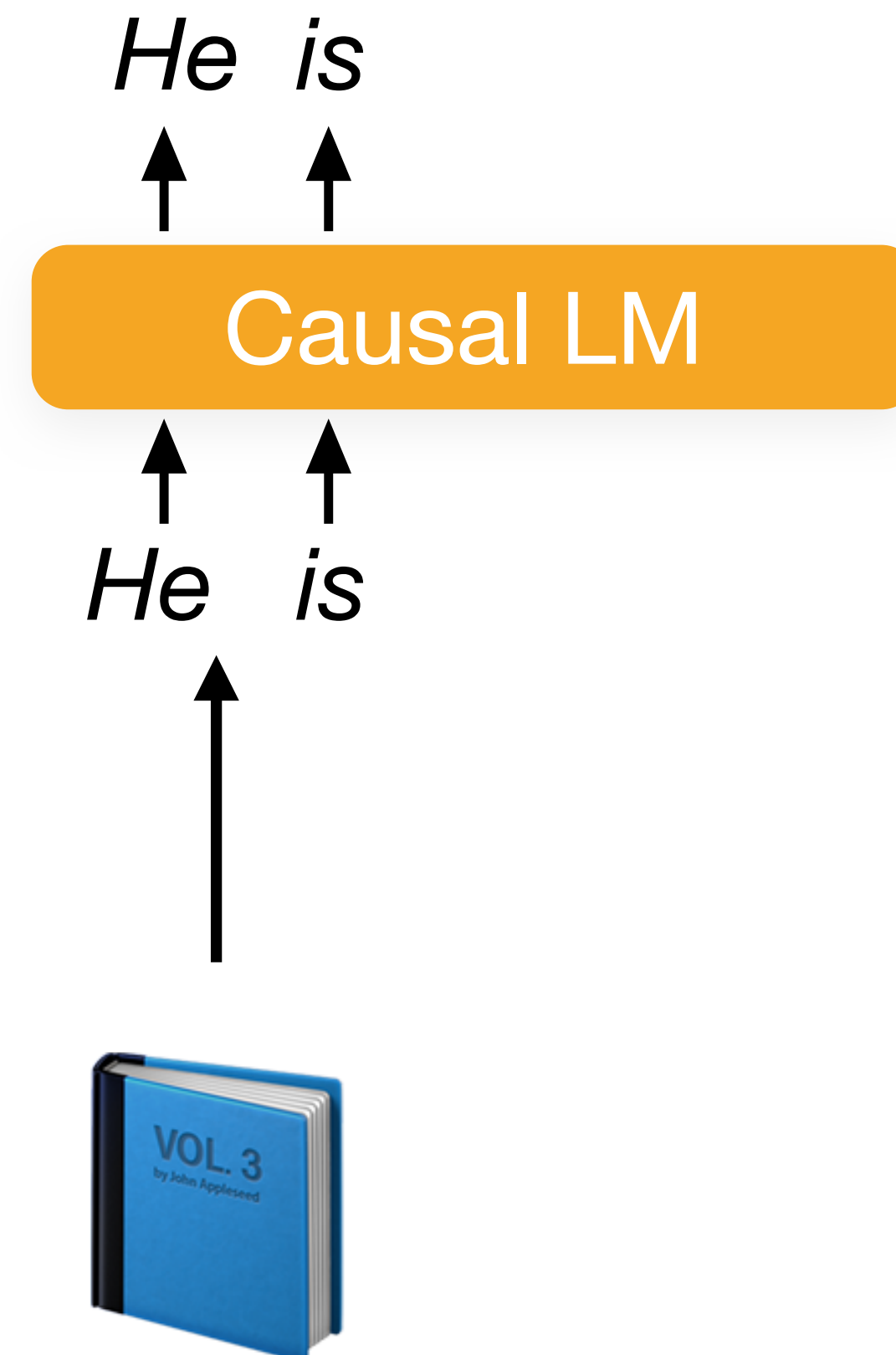
Each word affects the other words

Tokenizer

How a model understands text

Training data

What a model learns



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.

Models are created for English first

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

Models are created for English first

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.

Models are created for English first

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

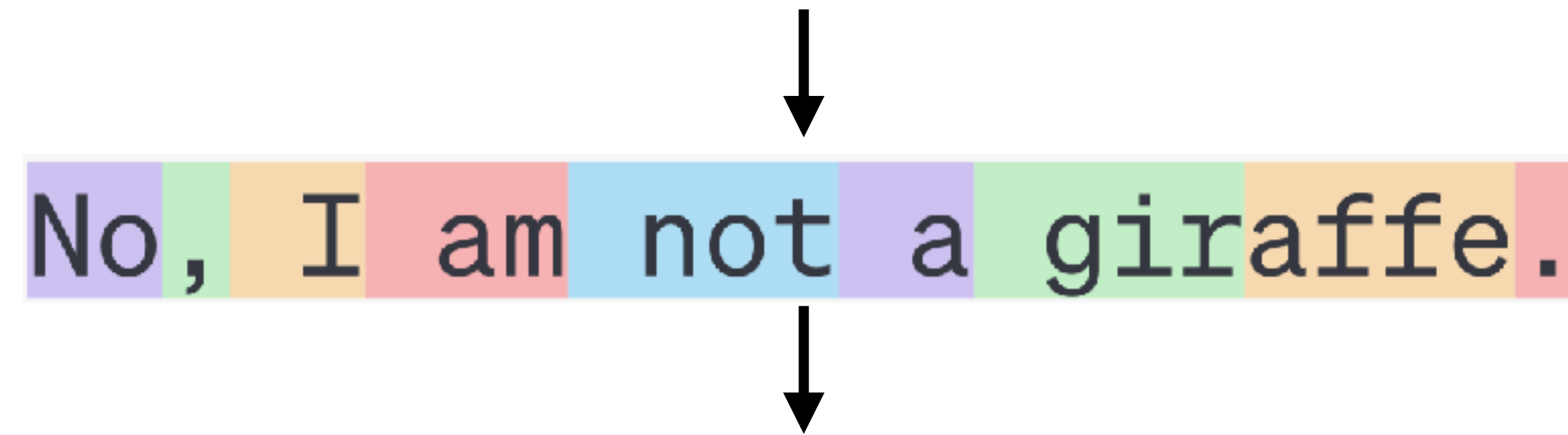
DE Nein, ich bin keine Giraffe. Das ist ein absurder Gedanke. → 1.50 x words

NL Nee, ik ben geen giraf. Dat is een absurde gedachte. → 1.50 x words

Tokenizing the training data

an example

No, I am not a giraffe.



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

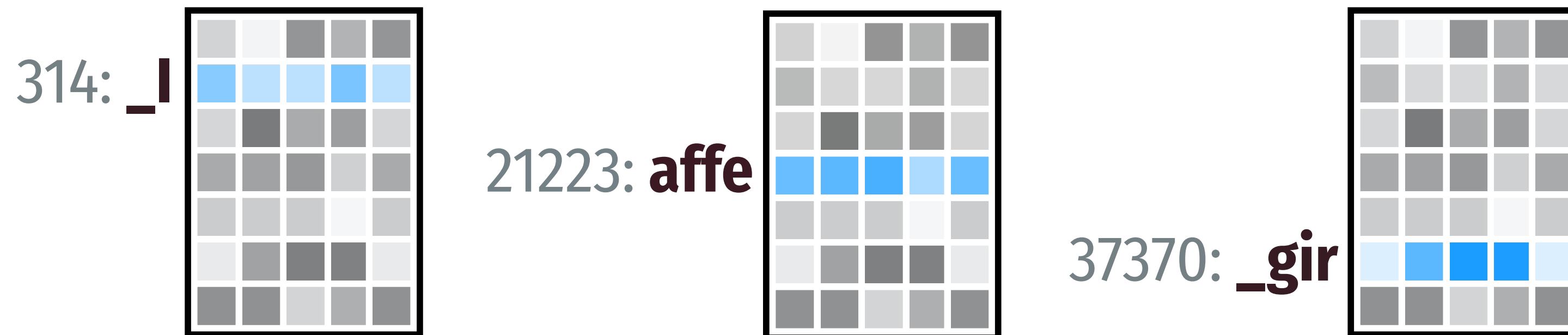
Tokenizing the training data

an example

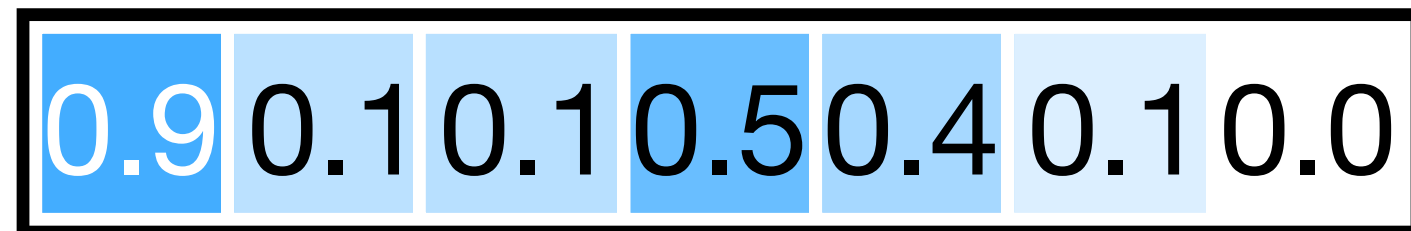
No, I am not a giraffe.

No, I am not a giraffe.

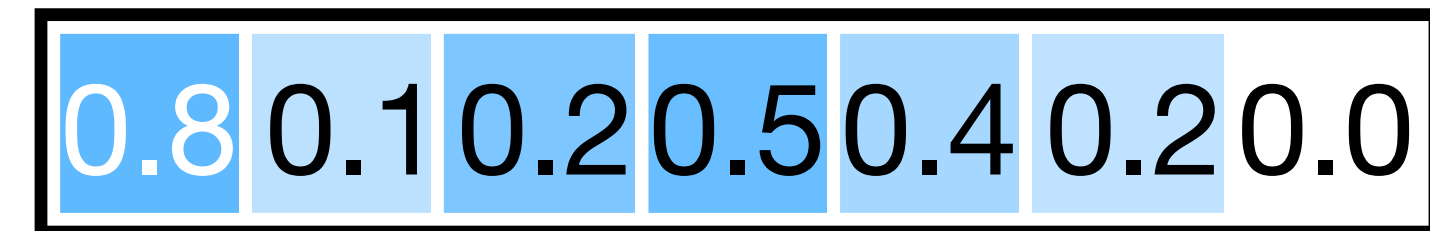
[2822, 11, 358, 1097, 539, 264, 37370, 21223, 13]



Embeddings capture meaning

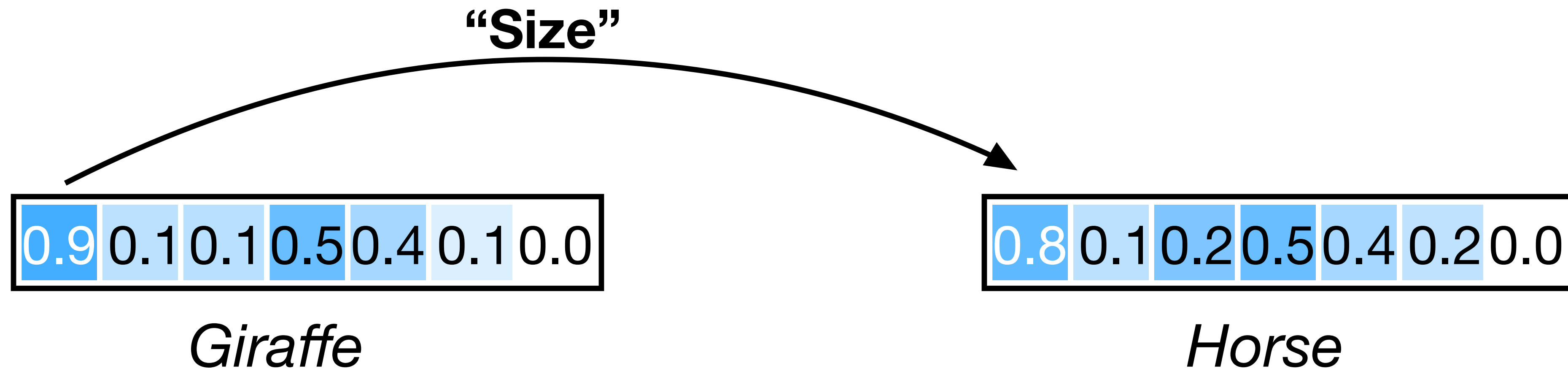


Giraffe

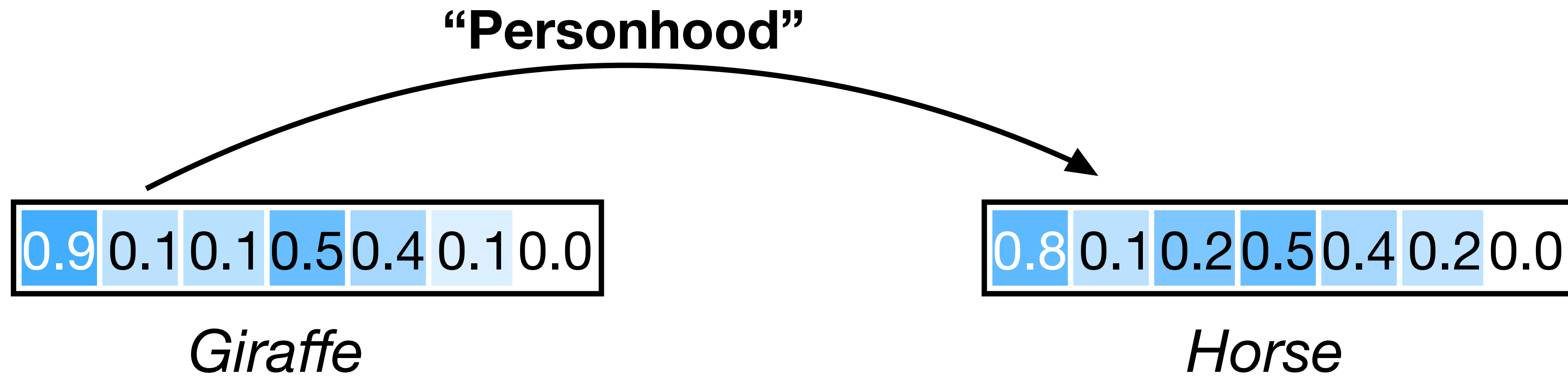


Horse

Embeddings capture meaning

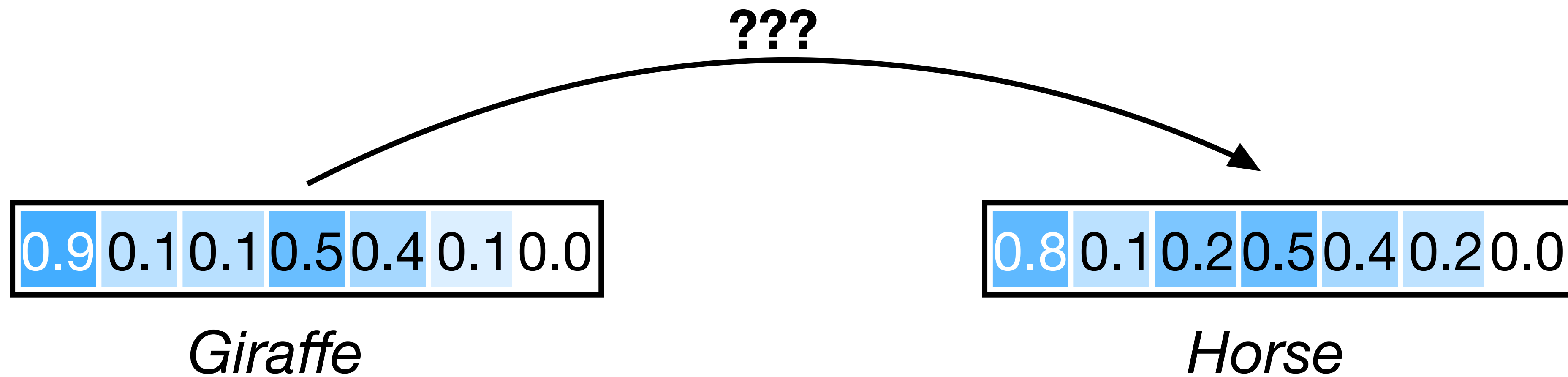


Embeddings capture meaning



Embeddings capture meaning

but it is not interpretable to us directly



Similar embeddings are close

Norway 

Sweden 

King 

Queen 


Giraffe 

Horse 

Similar embeddings are close

Norway 

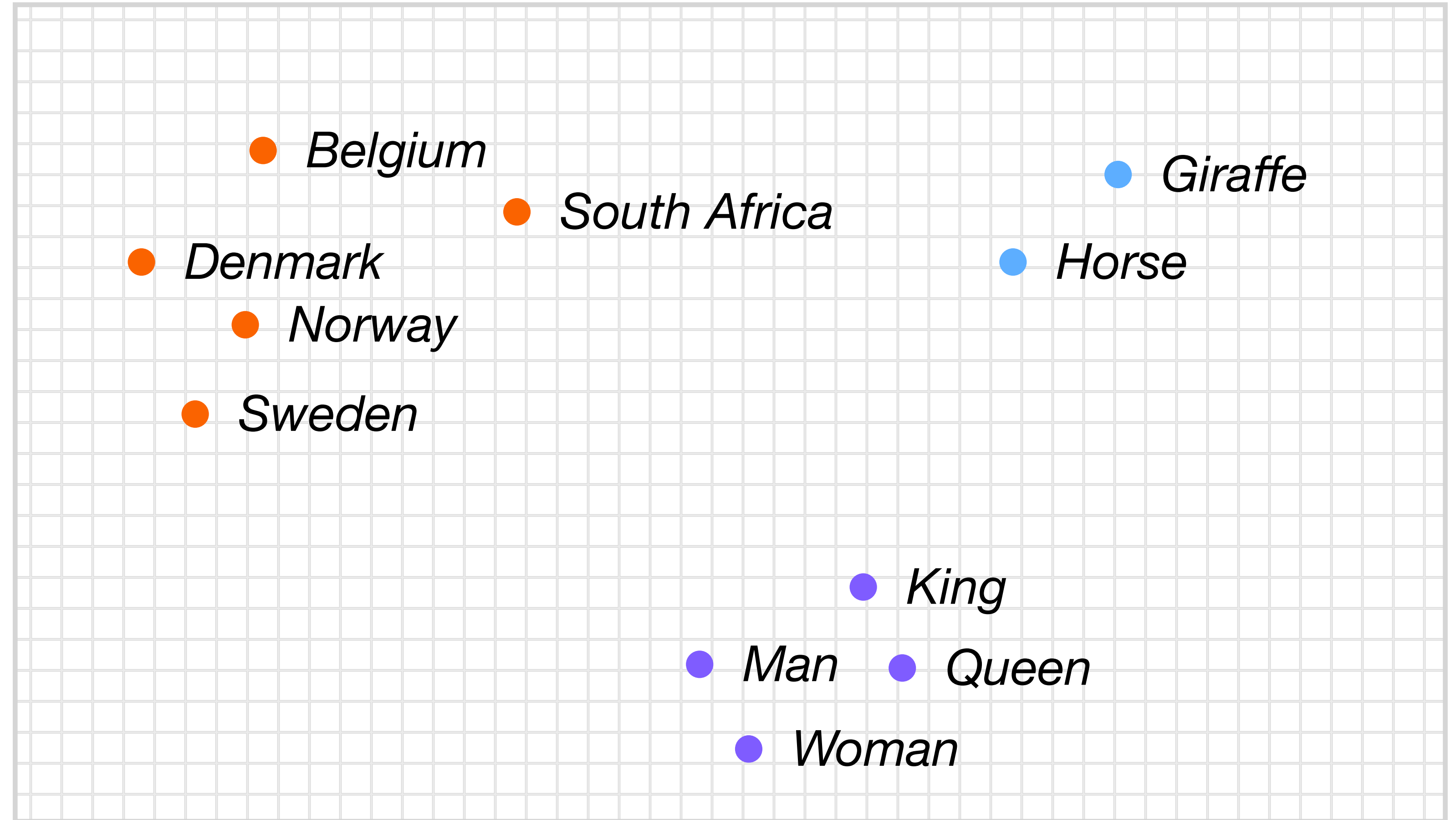
Sweden 

King 

Queen 

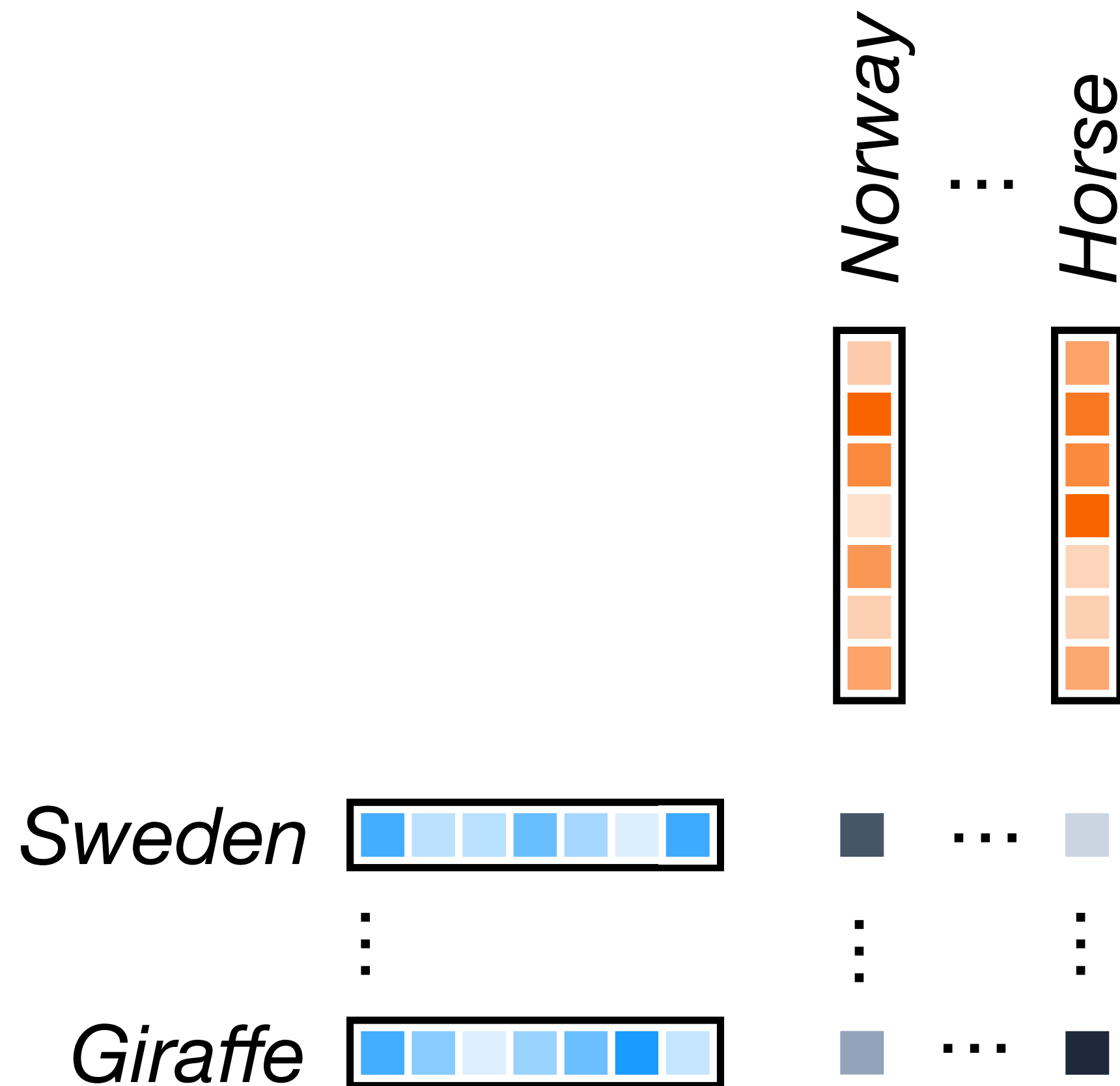
Giraffe 

Horse 



Similar embeddings are close

word embeddings



Word	Cosine distance
norway	0.760124
denmark	0.715460
finland	0.620022
switzerland	0.588132
belgium	0.585835
netherlands	0.574631
iceland	0.562368
estonia	0.547621
slovenia	0.531408

LLMs use context to learn embeddings

to address polysemy



Bank

LLMs use context to learn embeddings

to address polysemy



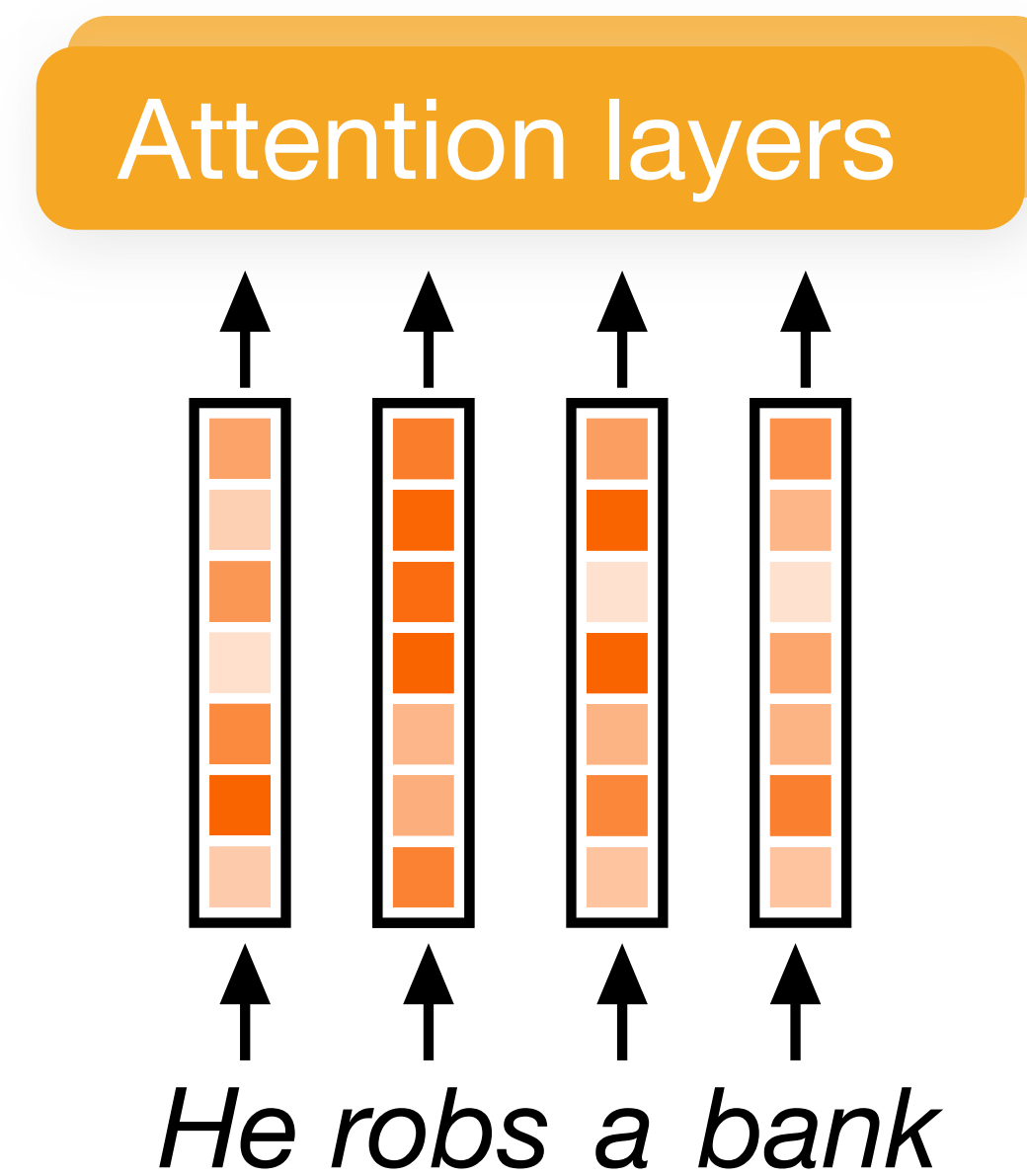
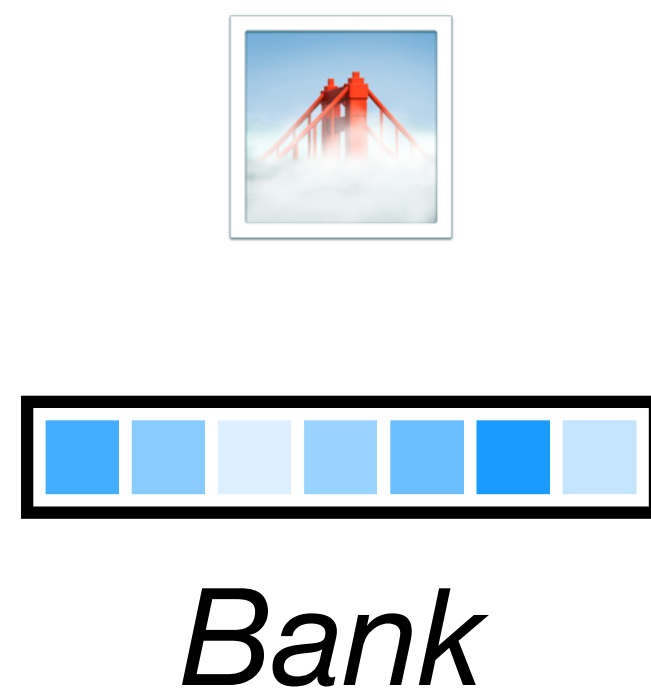
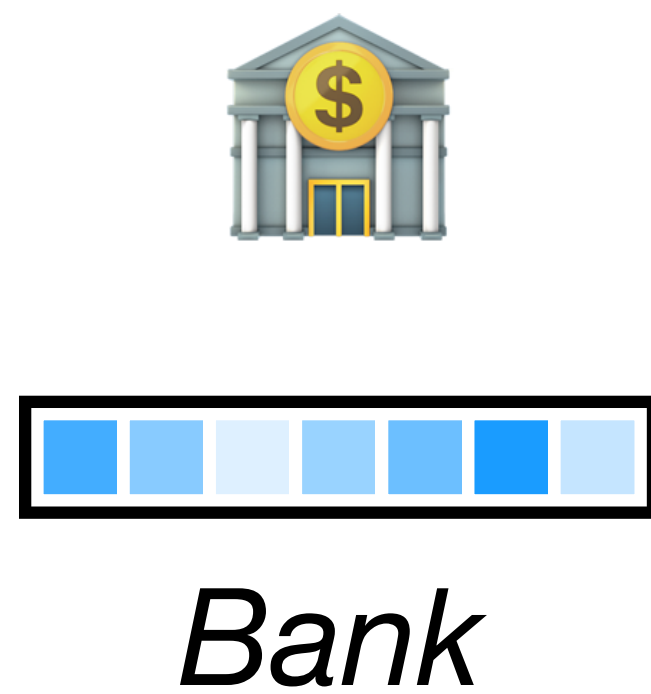
Bank



Bank

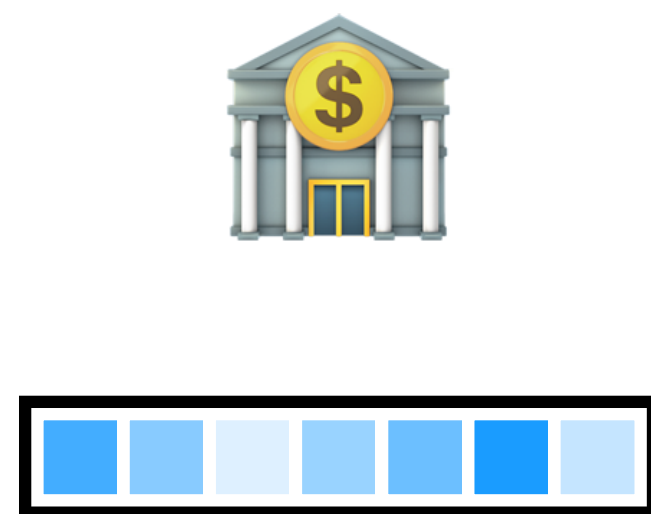
LLMs use context to learn embeddings

to address polysemy

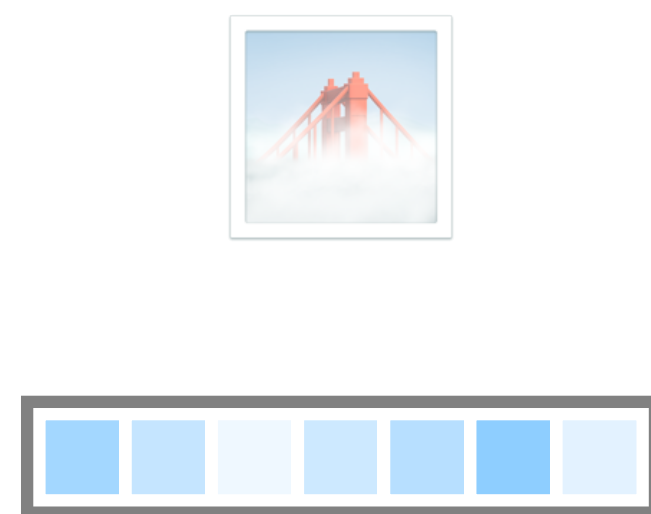


LLMs use context to learn embeddings

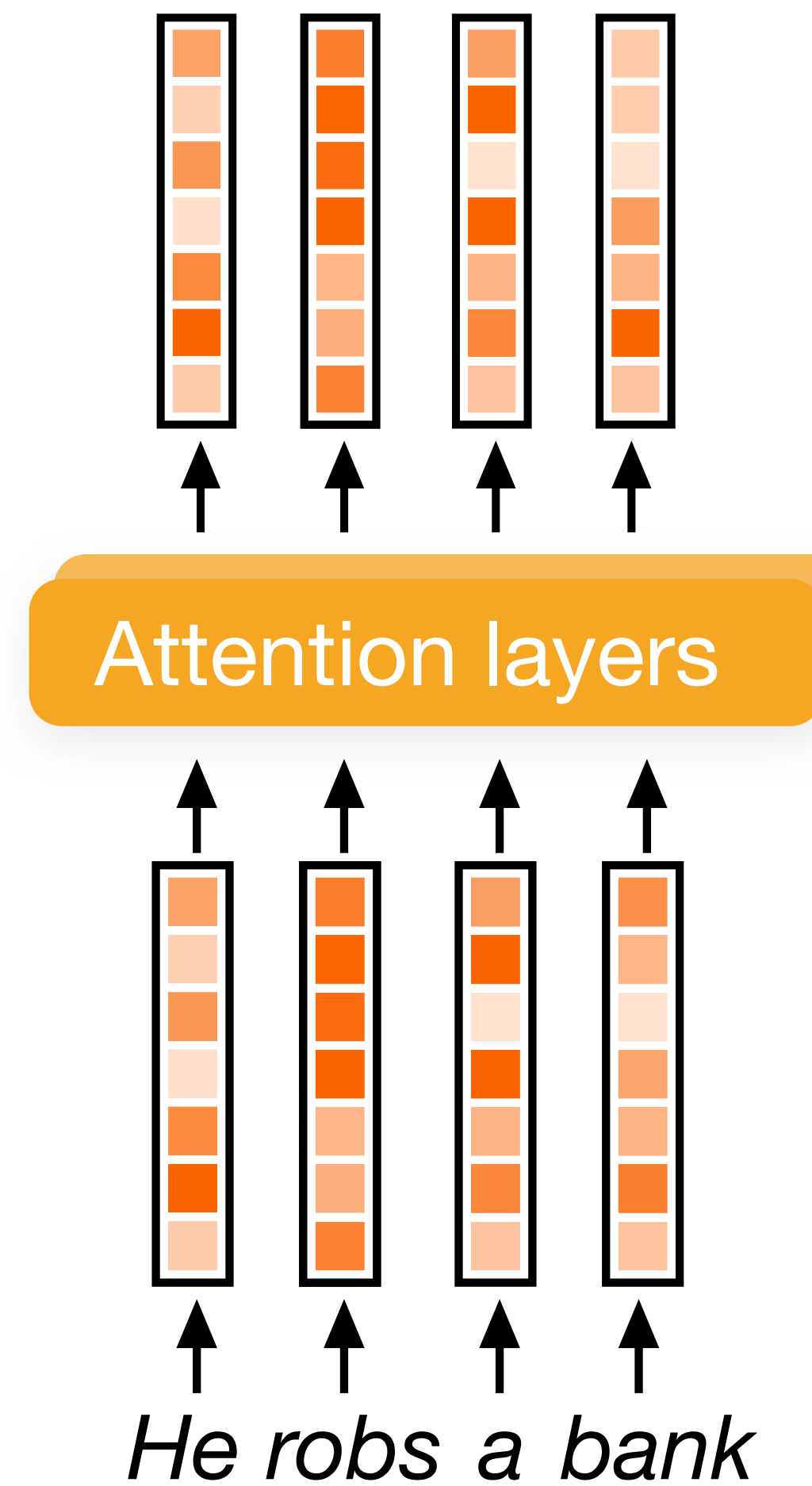
to address polysemy



Bank

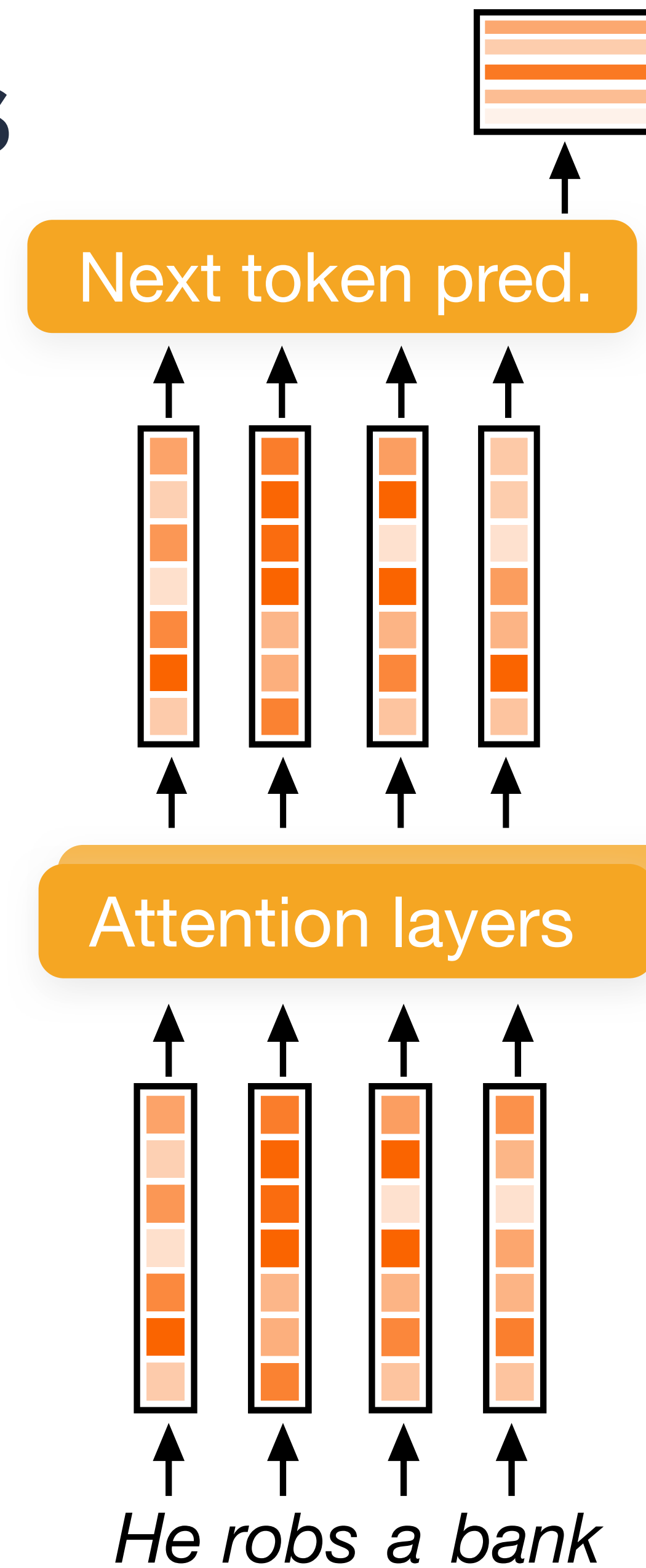


Bank



LLMs use context to learn embeddings

to address polysemy



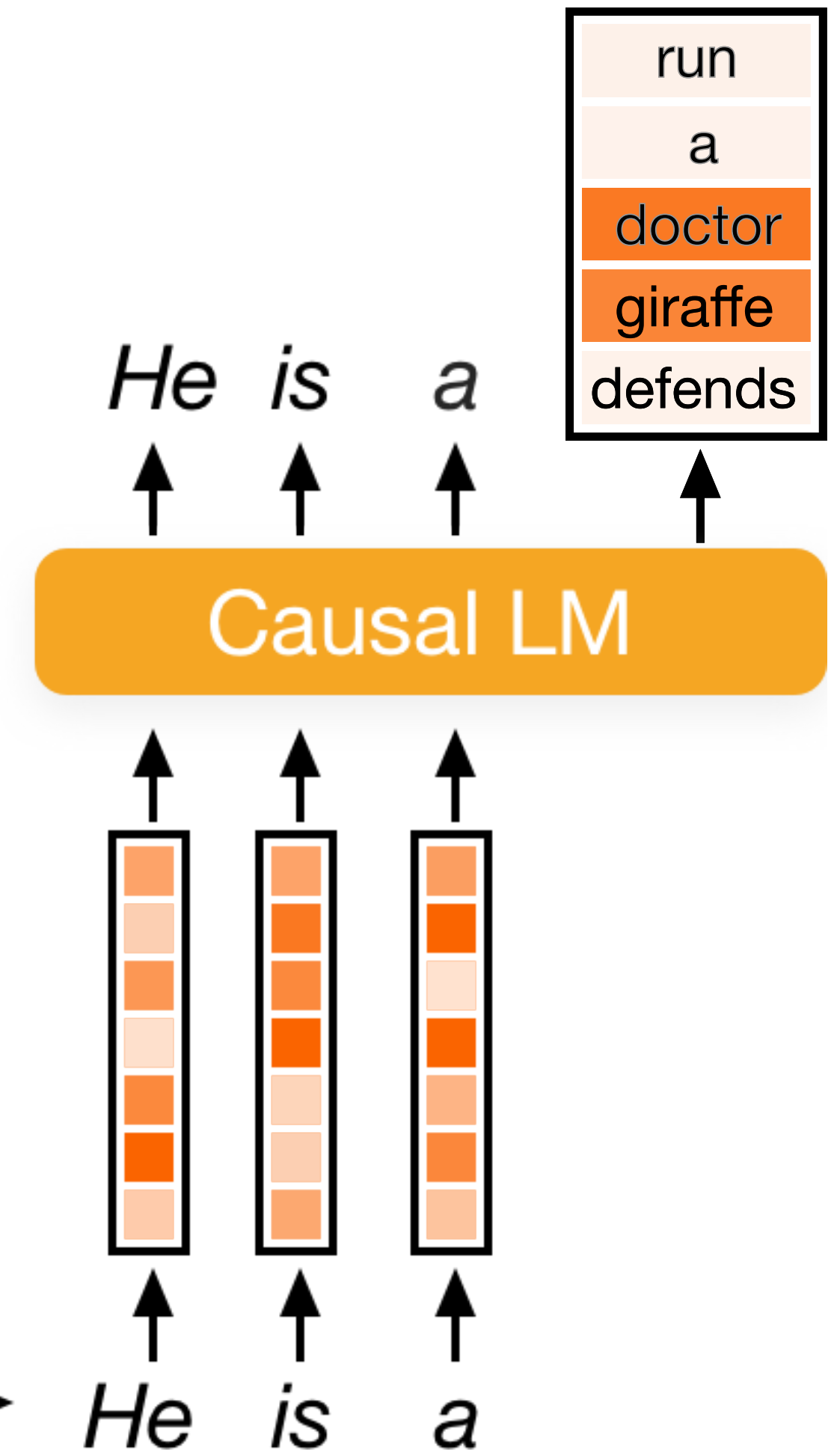
Predicting the next token

It is the tallest living terrestrial animal.

Giraffes live in herds.

He is a giraffe.

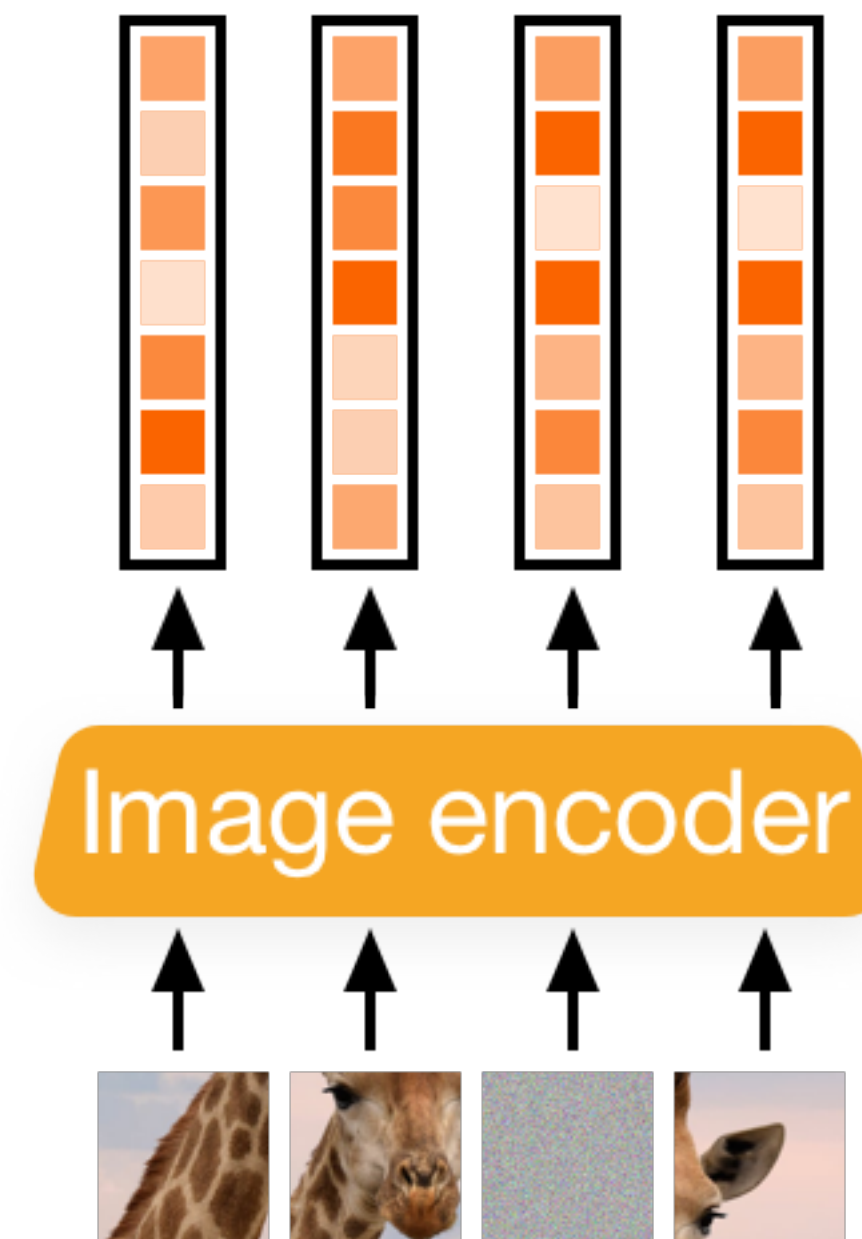
IUCN recognises one species of giraffe.



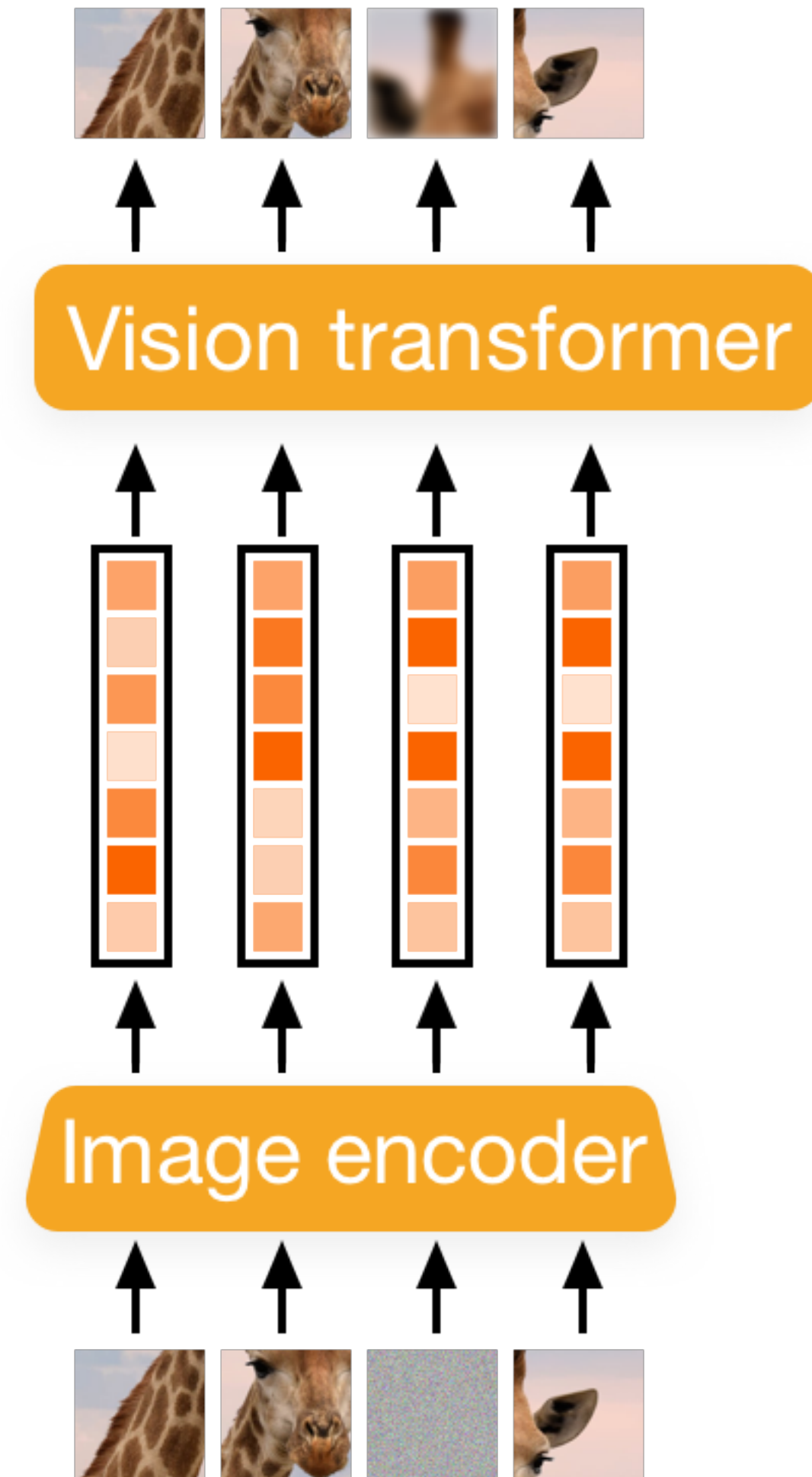
Multimodal language modeling



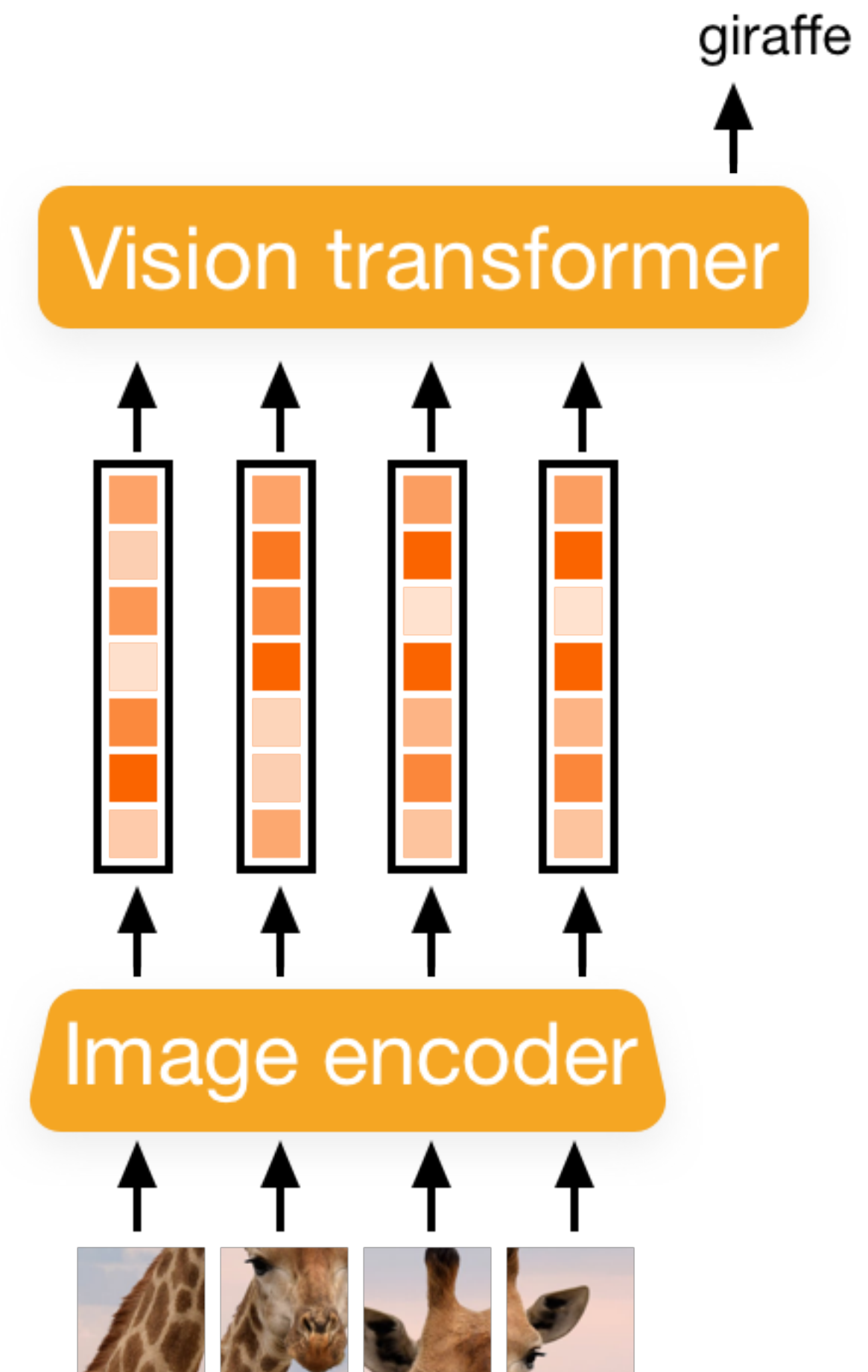
Multimodal language modeling



Multimodal language modeling



Multimodal language modeling



Instruction tuning

Teaching language models to follow conversations

Base model

Label the following sentence as positive or negative.

"I like giraffes."

Label:|

Positive

Label the following sentence as positive or negative.

"I like bananas

Instruction tuning

Teaching language models to follow conversations

Base model

Label the following sentence as positive or negative.

"I like giraffes."

Label:|

Positive

Label the following sentence as positive or negative.

"I like bananas

Instruction-tuned model with chat-templates

Label the following sentence as positive or negative. "I like giraffes."

Positive. The sentence expresses a liking or preference for giraffes.

```
<s>[INST] Label the following sentence as positive  
or negative... [/INST]"  
"Well, Positive. The sentence expresses a liking  
for ...</s> "  
"[INST] And this sentence: "... " [/INST]
```

Data security & sovereignty

Where does your data go?



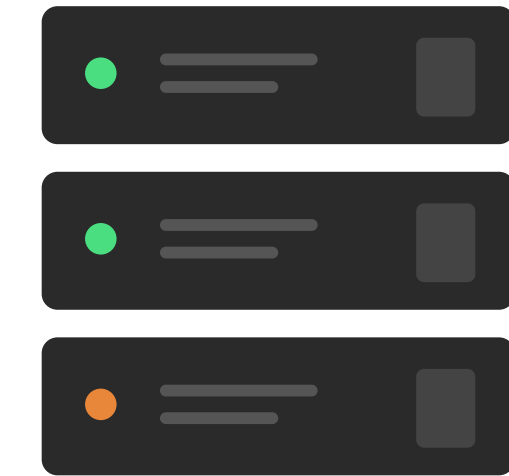
Your device

*prompts, documents,
client data, ...*

your prompt + context



generated response



Cloud servers

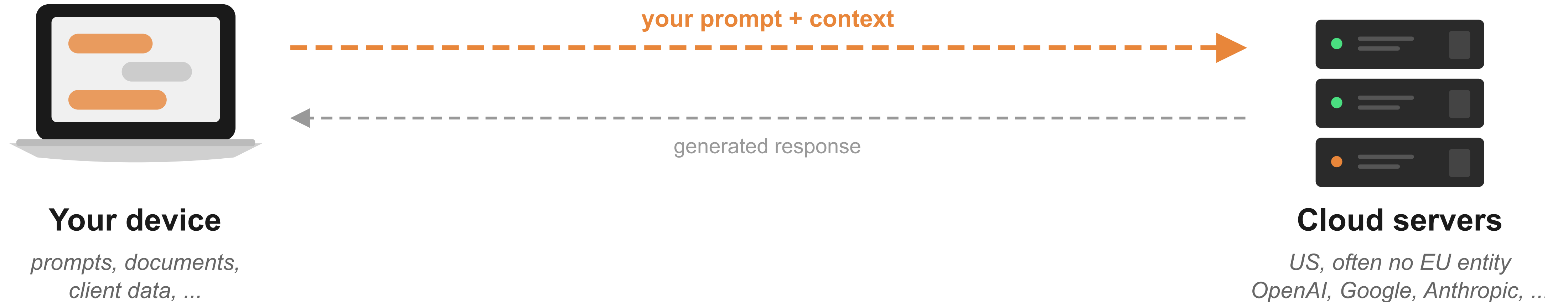
*US, often no EU entity
OpenAI, Google, Anthropic, ...*

Where does your data go?



1. Data *might* be used for training
2. Subject to US jurisdiction (CLOUD act)
3. No guarantee of confidentiality (*beroepsgeheim*)
4. GDPR compliance when dealing with personal data?

Where does your data go?



1. Data *might* be used for training → **Do not use free tiers**
2. Subject to US jurisdiction (CLOUD act)
3. No guarantee of confidentiality (*beroepsgeheim*)
4. GDPR compliance when dealing with personal data?

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 ET

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

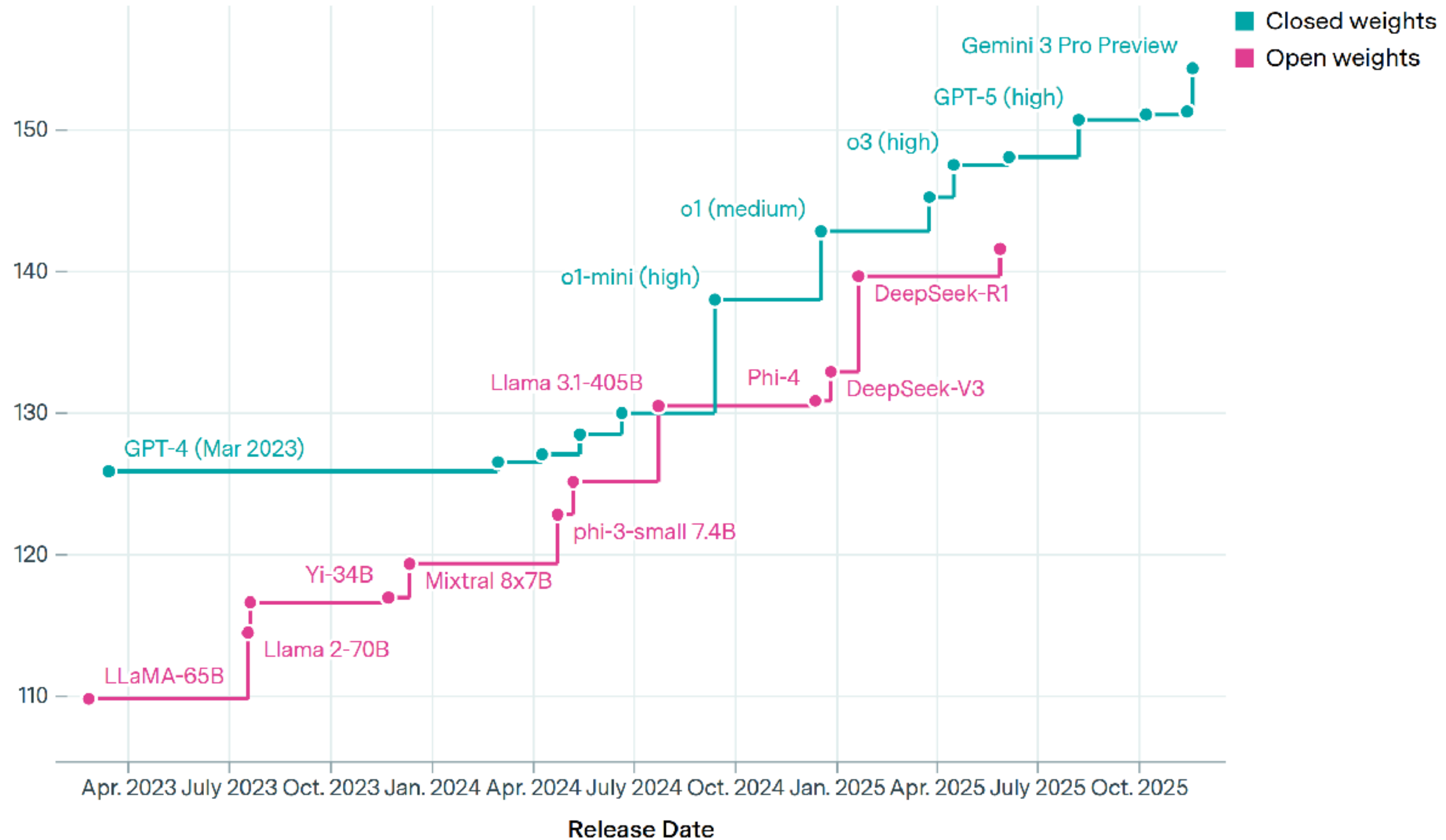
Published Jul 22, 2025, 07:56am EDT

→ so we want to run models “locally”

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

For EN

Epoch Capabilities Index score

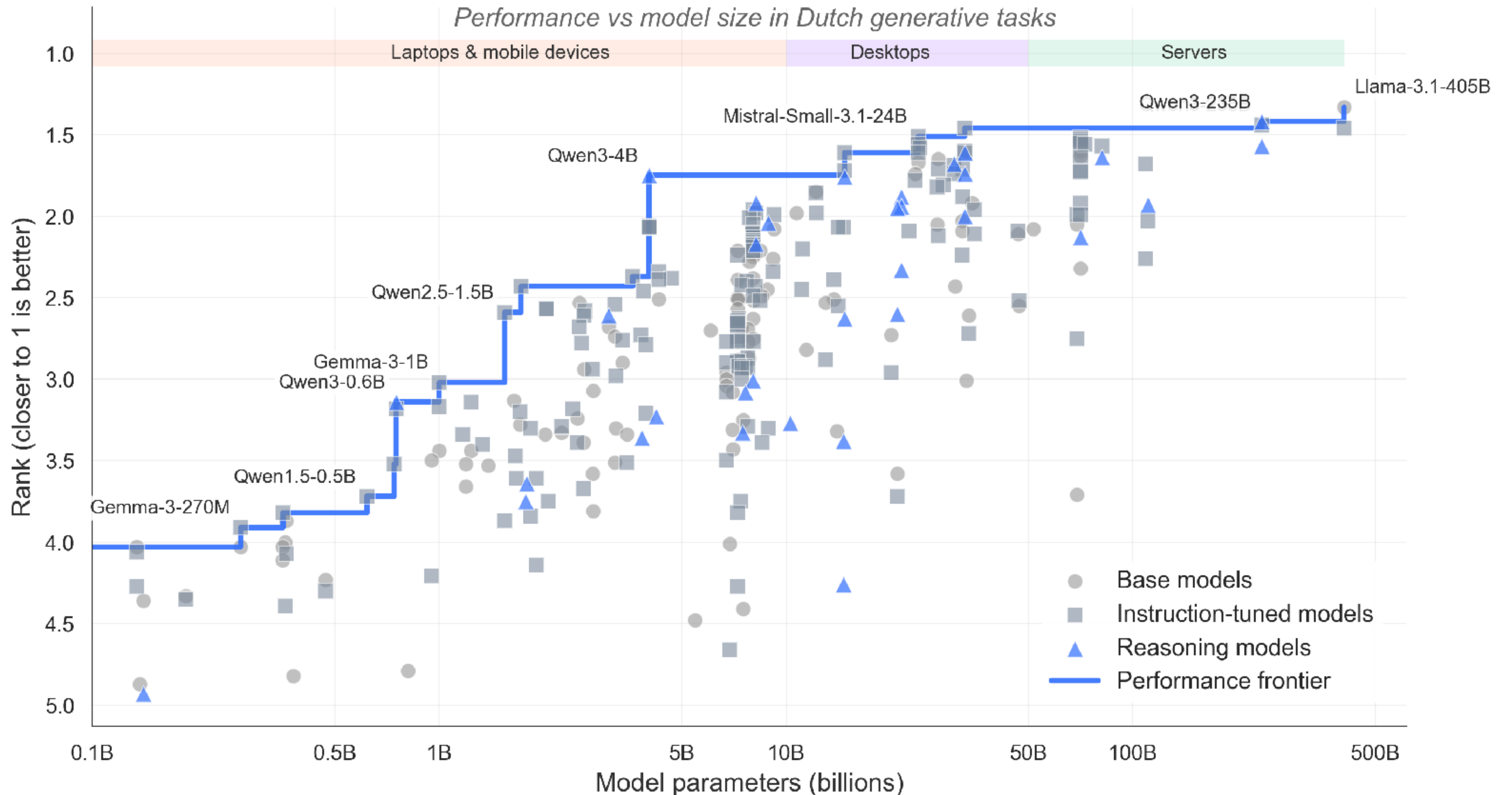


CC-BY

epoch.ai

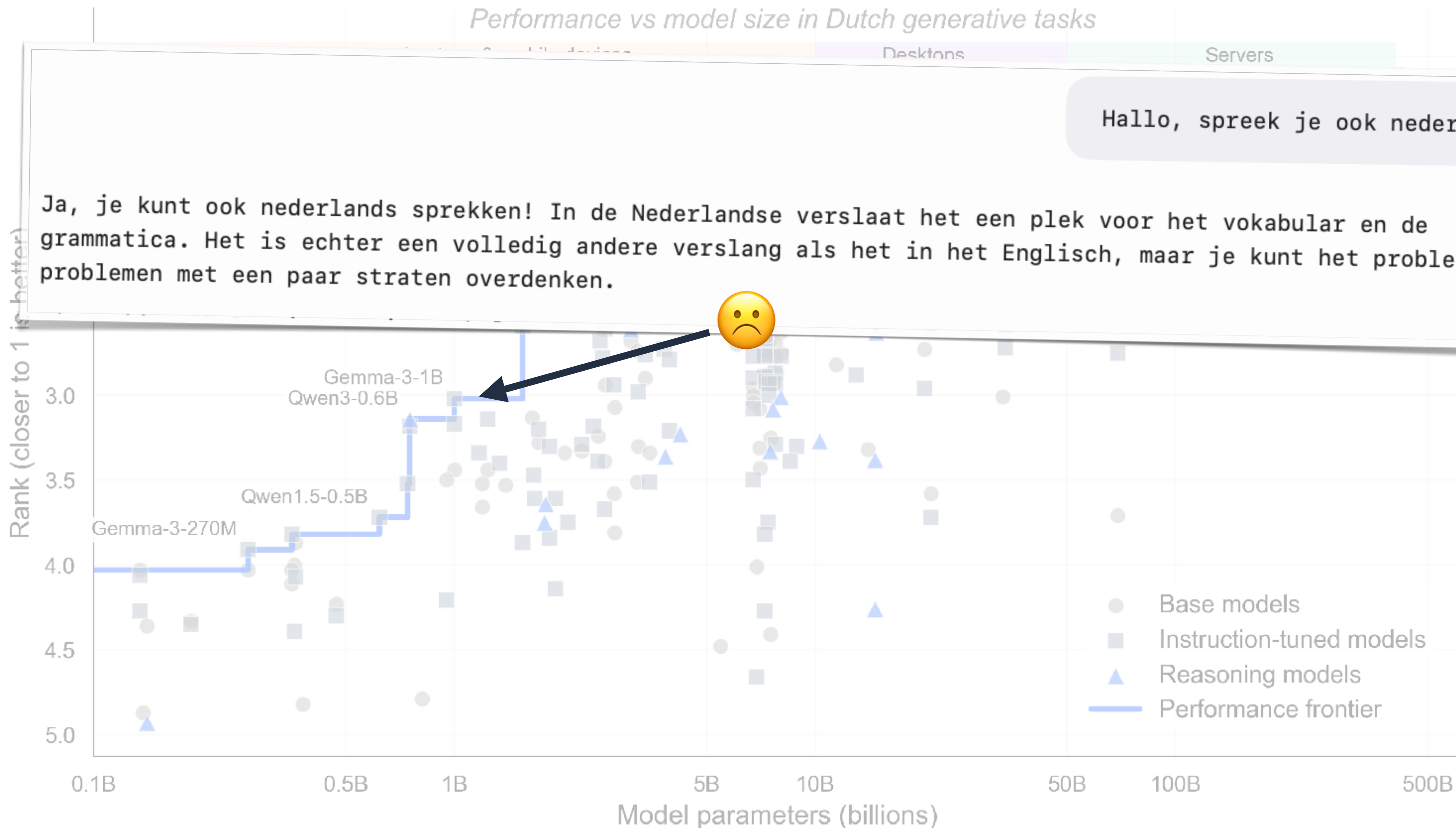
Larger models achieve better performance

For NL

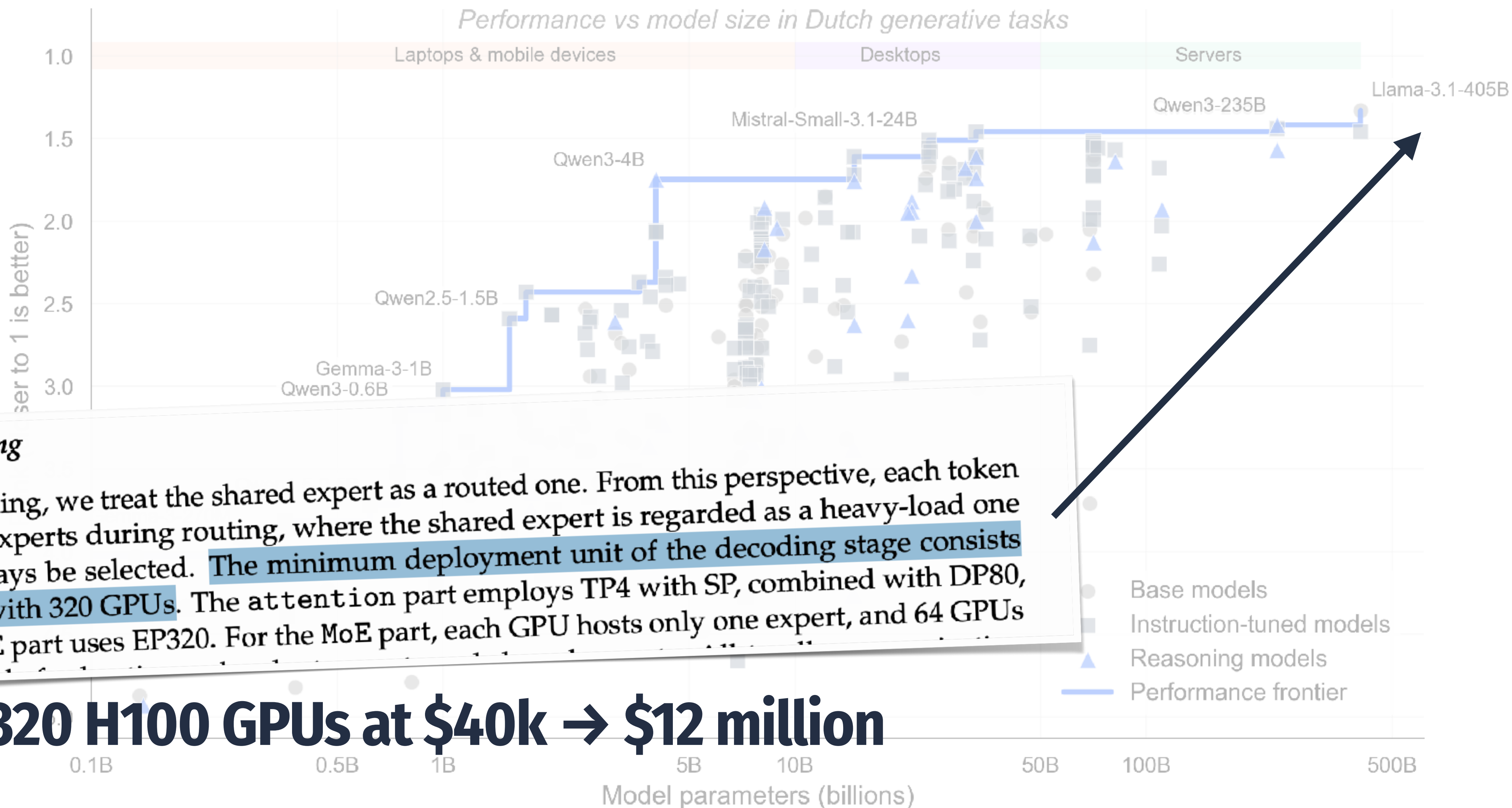


Source: euroeval.com/leaderboards/Monolingual/dutch | November 2025

Small models can run locally



... but deployment costs skyrocket for capable models



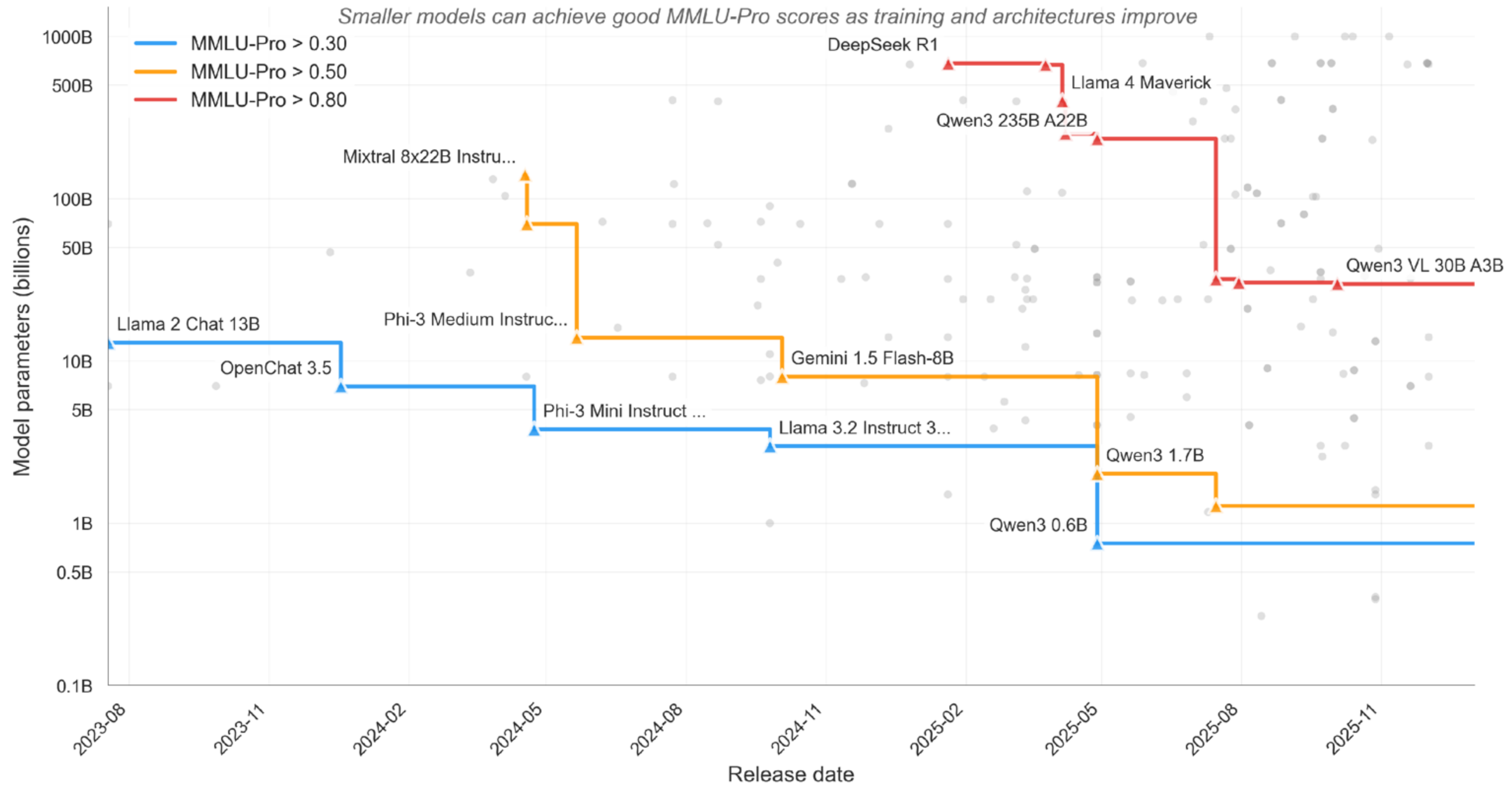
3.4.2. Decoding

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

Model quality

Smaller models are less capable but increasingly less so

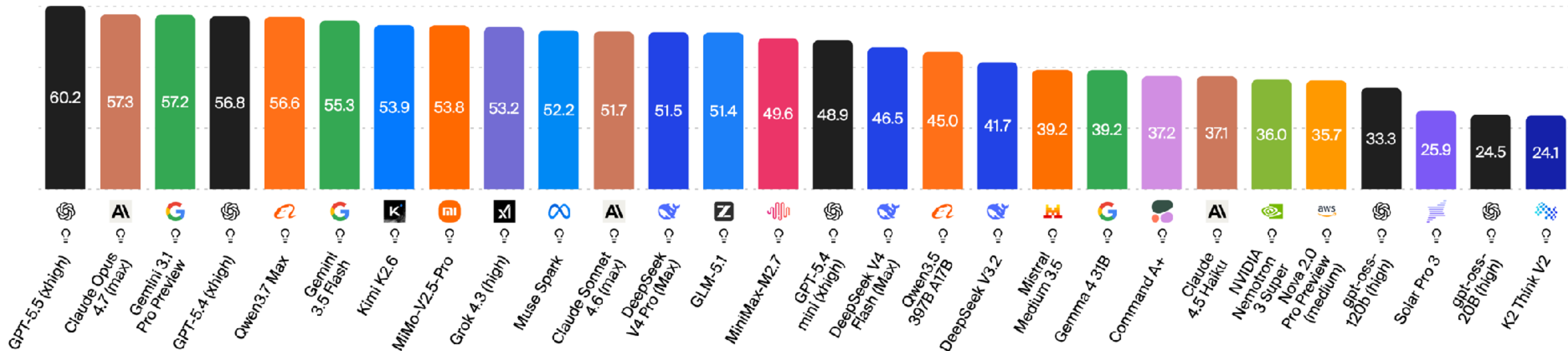


Frontier (closed) models are ahead

Artificial Analysis Intelligence Index

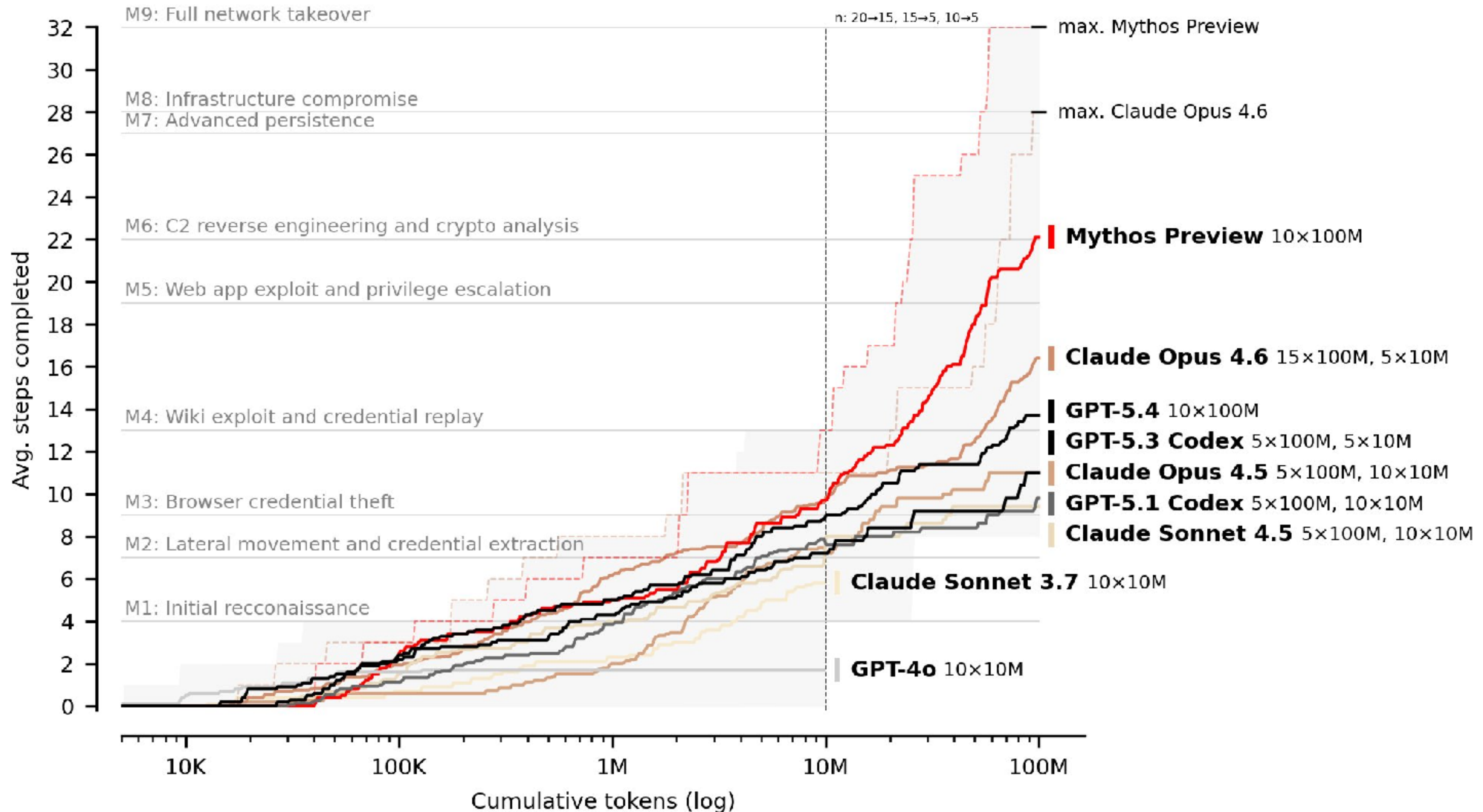
Artificial Analysis Intelligence Index v4.0 incorporates 10 evaluations: GDPval-AA, τ^2 -Bench Telecom, Terminal-Bench Hard, SciCode, AA-LCR, AA-Omniscience, IFBench, Humanity's Last Exam, GPQA Diamond, CritPt

Artificial Analysis



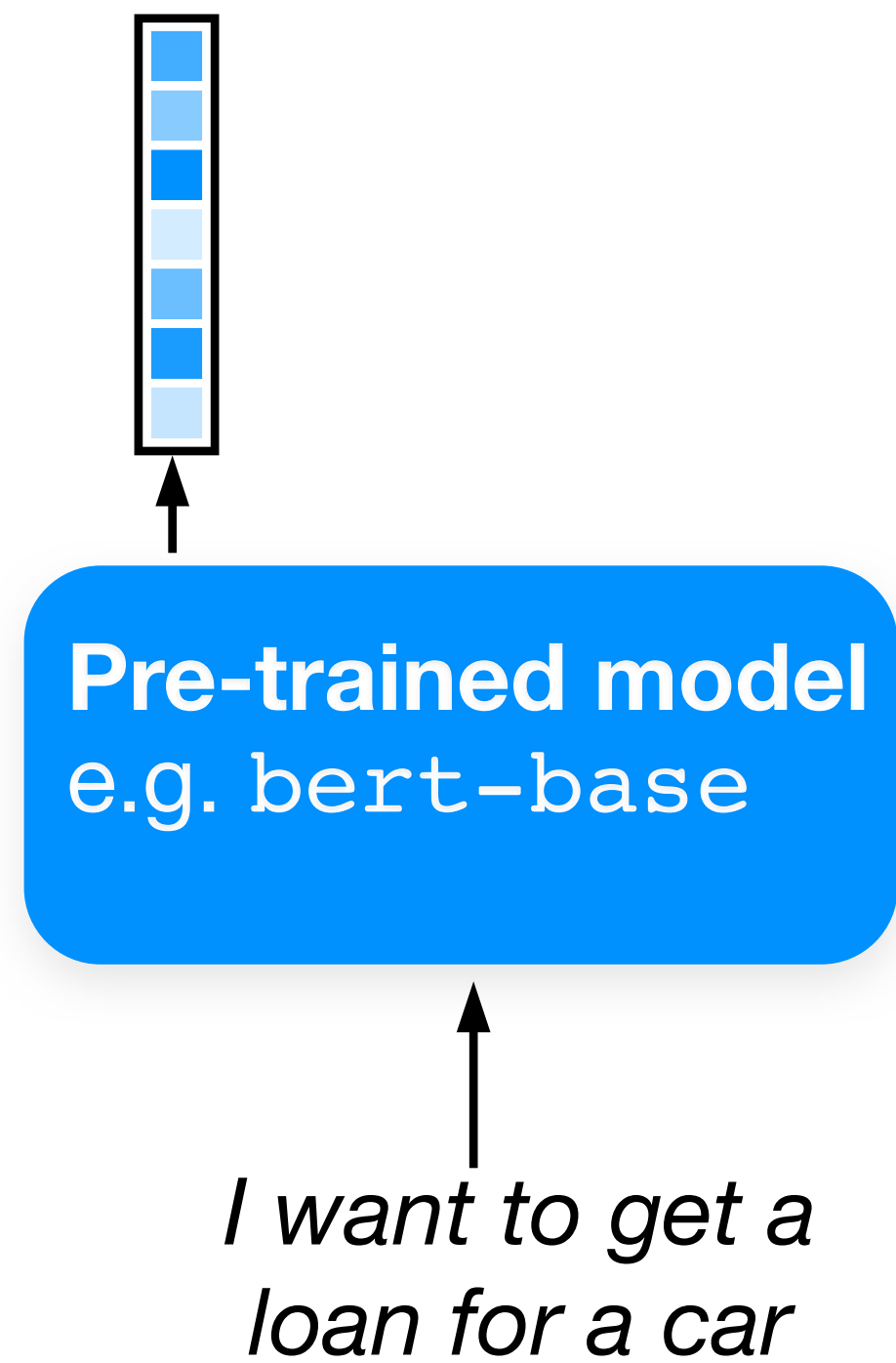
Reasoning models are indicated by a lightbulb icon

Frontier (closed) models are ahead especially for coding

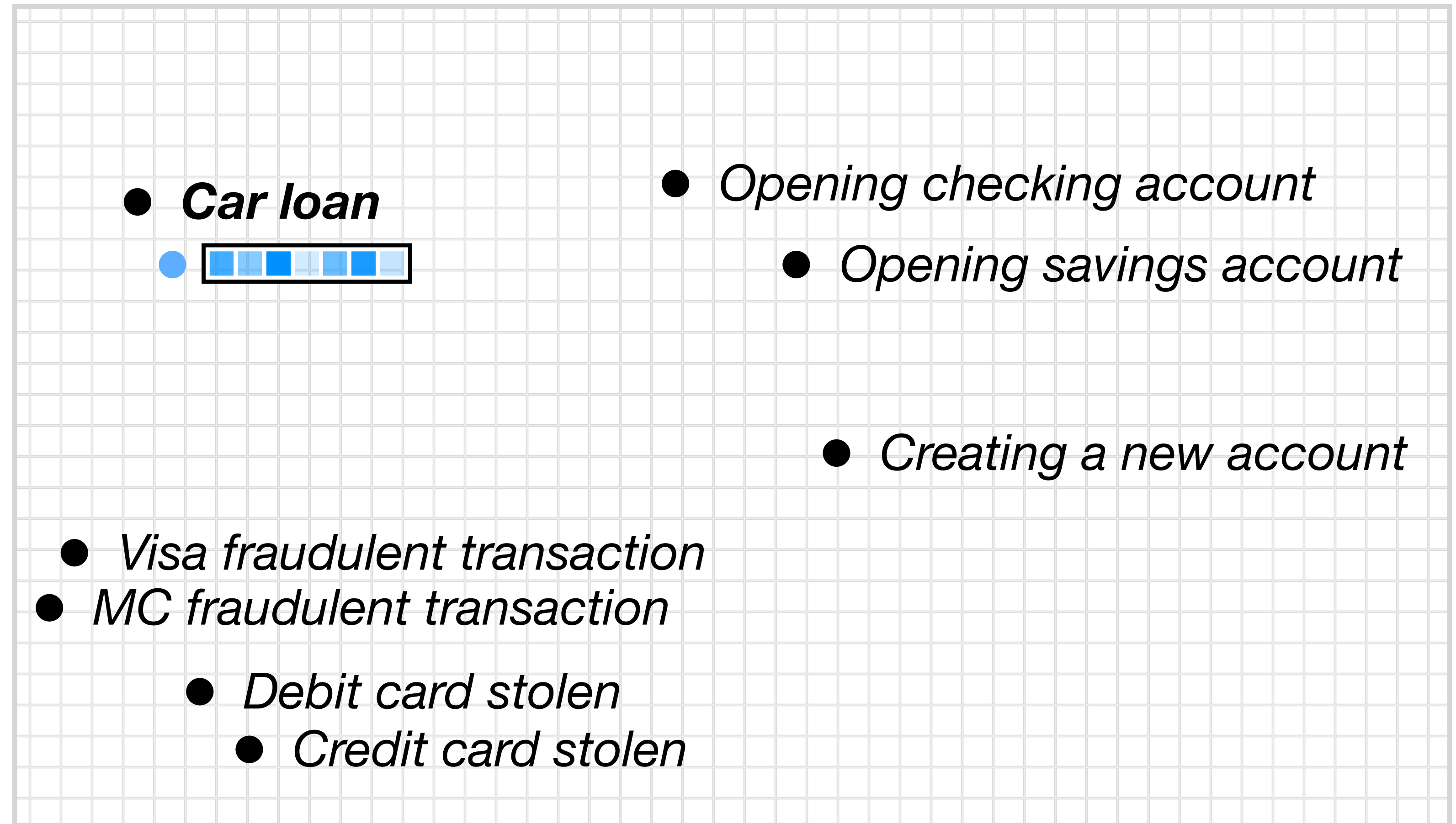
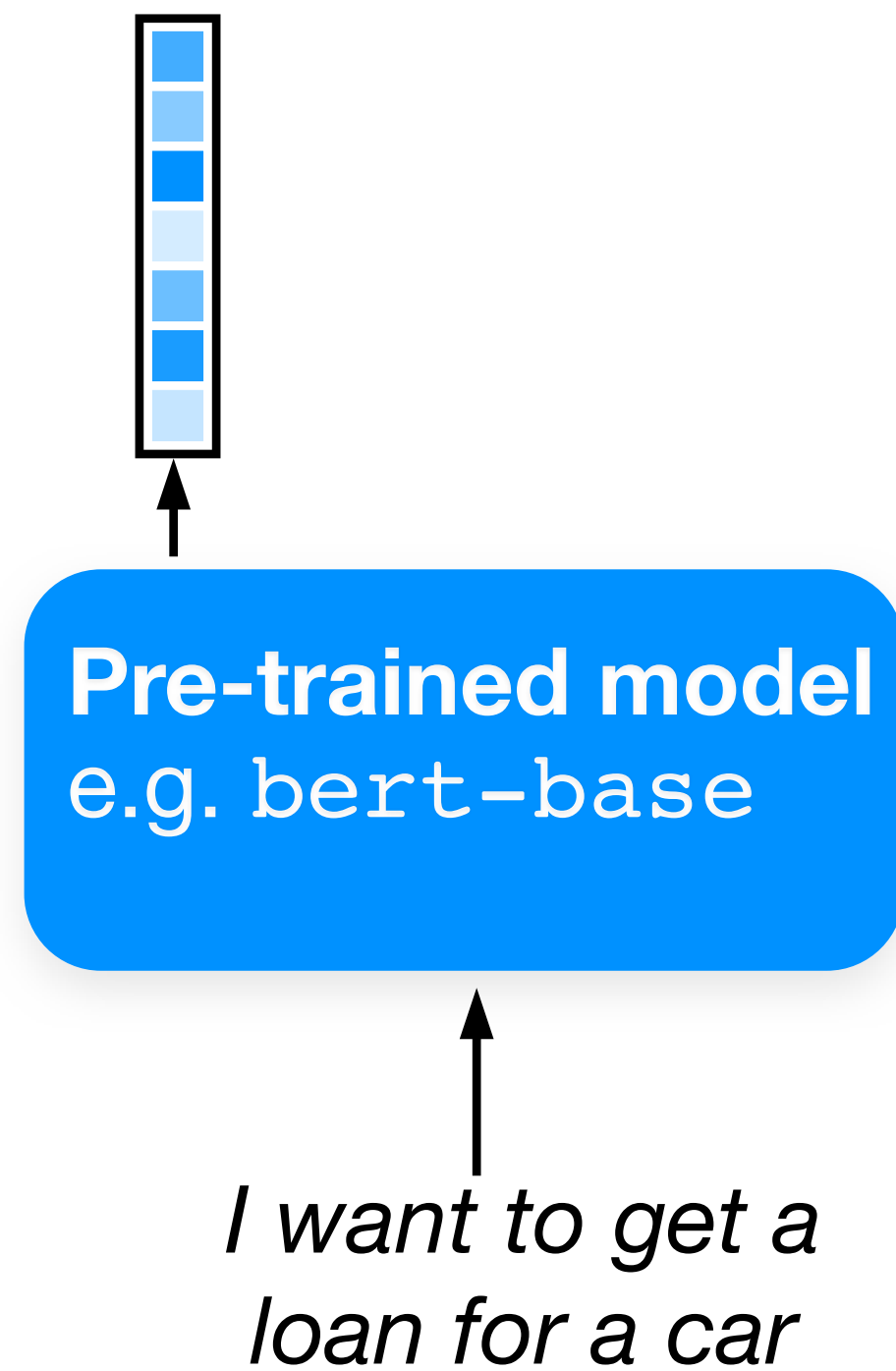


Hallucinations and retrieval augmented generation

Semantic search



Semantic search



Retrieval augmented generation (RAG)

Default

Oudenaarde is een historisch stadje in de provincie **West-Vlaanderen** in België. Het ligt in het oosten van de Vlaanderen, aan de **Zenne**. De stad is gelegen op een strategische locatie **tussen** Brussel en Gent. De geschiedenis van Oudenaarde ontwikkelde zich vanwege haar strategische ligging als een vestingplaats. Al in de **9e eeuw was de stad de zetel van de Vikingen in Zeeland. In de tweede helft van de 10e eeuw wordt het opgericht als de zetel van de Grieken en later van de Normannen.** In de 14e eeuw komt de stad **onder de Zuidelijke Nederlanden, toen wordt het stadsrecht het bestuur over de stad verlenen.**

De stad heeft veel historische bouwwerken, zoals het stadhuis, de **Korenlei en het Klokhuis.** Het **historisch centrum** wordt beschermd als UNESCO Werelderfgoed.

Oudenaarde is tevens de **hoofdplaats van de provincie Oost-Vlaanderen.** Het stadje is bekend om zijn laken, de traditionele wederzijdsstrijd tussen Oudenaarde en Gent, en de traditionele feestdag van **6 augustus,** toen de **Slag van de Drie Heeren** gelegen is.

RAG

Oudenaarde is een stad in de provincie Oost-Vlaanderen in België. Het is de tweede cultuurstad van de provincie na Gent en telt veel historische gebouwen. De stad is bekend om haar bruine bieren, zoals Cnudde, Ename, Felix, Liefmans en Roman. Oudenaarde heeft een rijk cultureel erfgoed en is bekend om haar landelijke omgeving in de Vlaamse Ardennen. De stad heeft een belangrijke toeristische industrie. Oudenaarde is ook bekend om haar textielnijverheid, zoals de textielfabriek Saffre Frères.



AI-generated fake “facts”: **hallucinations**

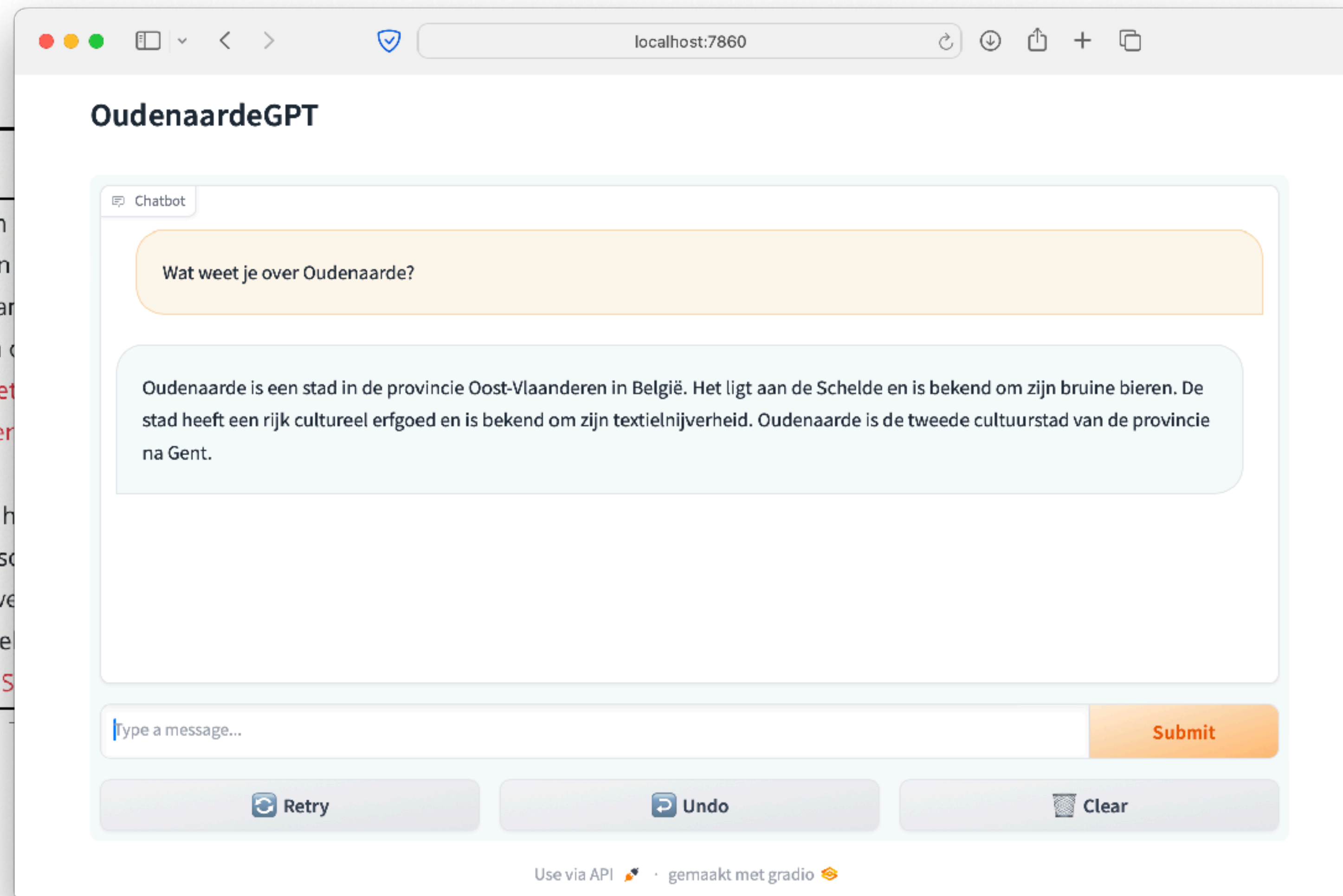
Retrieval augmented generation (RAG)

Default

Oudenaarde is een stad in de provincie Oost-Vlaanderen, aan de Schelde. De geschiedenis van de stad gaat terug tot de vestingplaats. Al in de 10e eeuw wordt het bekend dat de stad onder de Franse verlenen.

De stad heeft veel historische gebouwen. Het centrum wordt beschouwd als een van de mooiste van België. Oudenaarde is tevens bekend om zijn laken, de traditionele textielindustrie. In augustus, toen de Sint-Augustusfeesten worden gevierd.

De provincie Oost-Vlaanderen heeft een rijk cultureel erfgoed en is bekend om zijn historische steden. Oudenaarde is de tweede cultuurstad van de provincie na Gent. Het heeft een prachtige omgeving met veel historische gebouwen. Oudenaarde is bekend om zijn laken, de traditionele textielindustrie. In augustus, toen de Sint-Augustusfeesten worden gevierd.



... but not a silver bullet



Lawyer apologizes for fake court citations from ChatGPT

By [Ramishah Maruf](#), CNN

The New York Times

Here's What Happens When Your Lawyer Uses ChatGPT

A lawyer representing a man who sued an airline relied on artificial intelligence to help prepare a court filing. It did not go well.

... but not a silver bullet



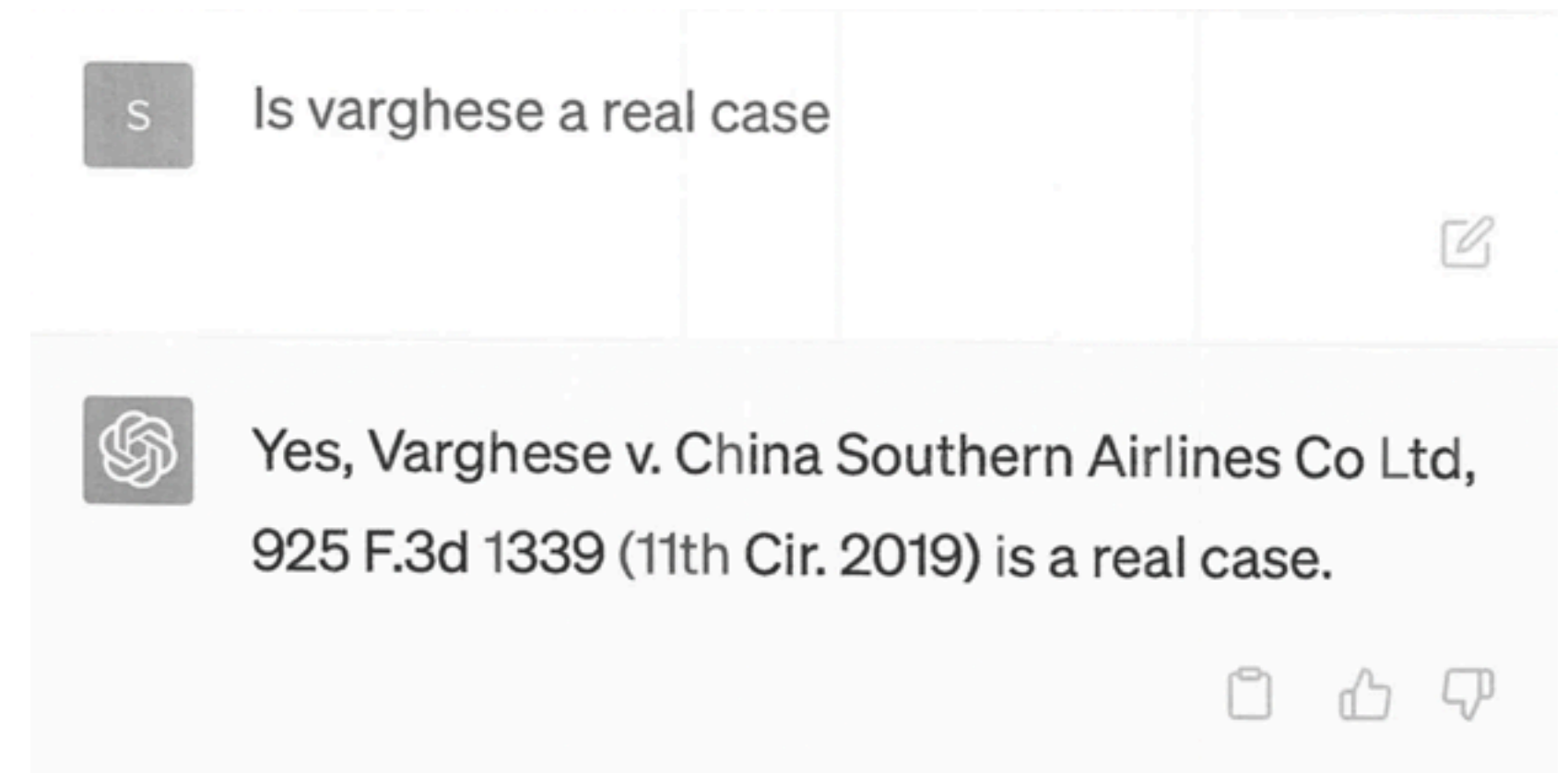
Lawyer apologizes for fake court citations from ChatGPT

By [Ramishah Maruf](#), CNN

The New York Times

Here's What Happens When Your Lawyer Uses ChatGPT

A lawyer representing a man who sued an airline relied on artificial intelligence to help prepare a court filing. It did not go well.



... but not a silver bullet



Lawyer apologizes for fake court citation ChatGPT

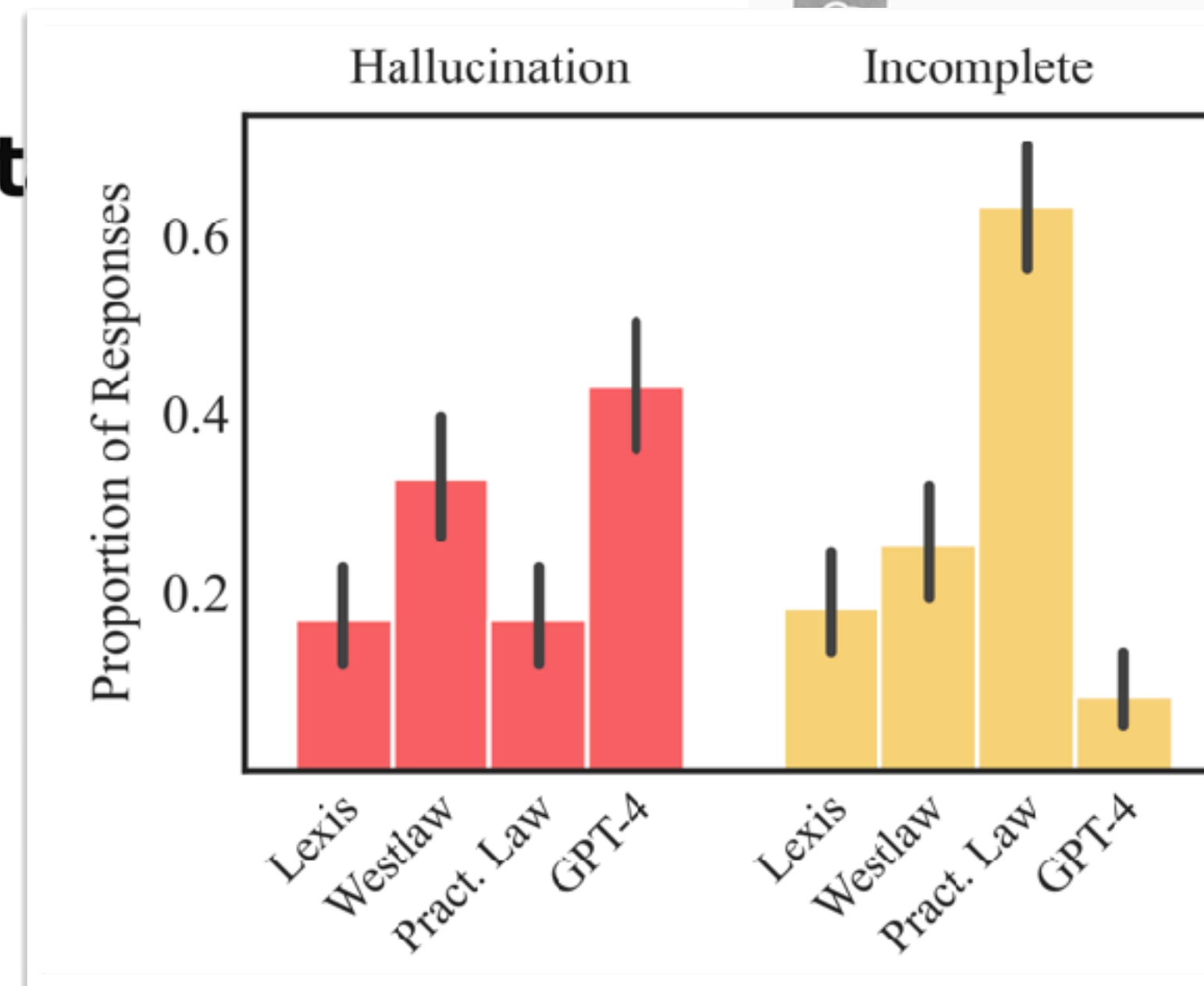
By Ramishah Maruf, CNN

The New York Times

Here's What Happens When Your Lawyer Uses ChatGPT

A lawyer representing a man who sued an airline relied on artificial intelligence to help prepare a court filing. It did not go well.

→ *Hallucinations* are inherent to LLMs



Is varghese a real case

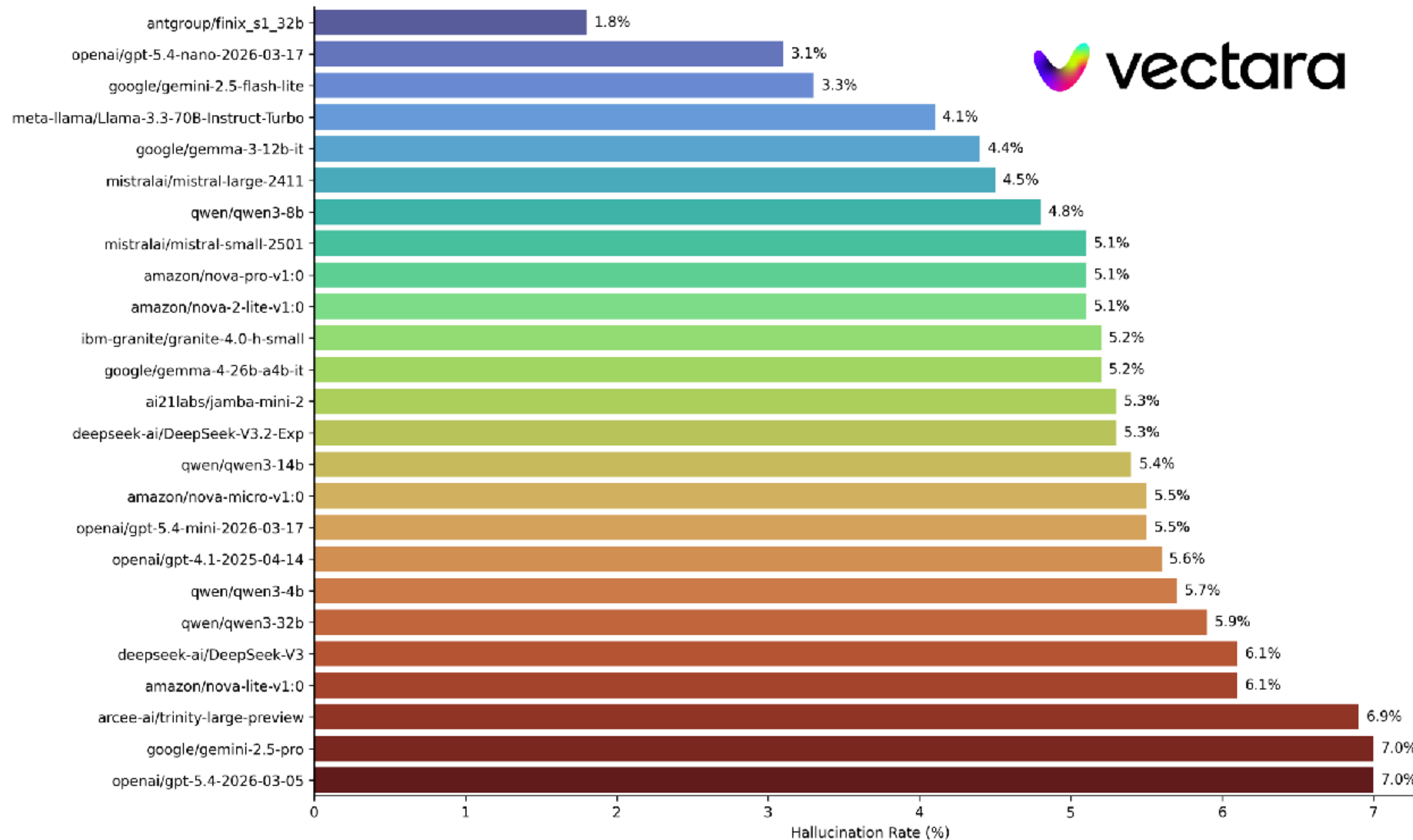
China Southern Airlines Co Ltd, (925 F.3d 1339, 13th Cir. 2019) is a real case.

... confusion earlier. Upon further research I found that the case Varghese Airlines Co. Ltd., 925 F.3d 1339 indeed exist and can be found on legal research databases such as Westlaw and LexisNexis. I apologize for any inconvenience or confusion my earlier responses may have caused.

Modern models do not hallucinate a lot

But not 0%

Grounded Hallucination Rates for Top 25 LLMs



Models might lack context

kan ik gerechtskosten terug krijgen als ik de rechtzaak win?

Models might lack context

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

Models might lack context

kan ik gerechtskosten terug krijgen als ik de rechtzaak win?

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

Kostenveroordeling (artikel 1019h Rv)

Wanneer je als winnende partij optreedt, kan de rechtbank de betaling van je gerechtskosten. Dit gebeurt meestal automatisch in het vonnis, tenzij anderszins bepaald (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



Wetboek Plus

<https://wetboekplus.nl> > burgerlijke-rechtsvordering-boek-Derde-artikel-1019h

...

Artikel 1019h Rv - Wetboek+

Voor zover nodig in afwijking van de tweede paragraaf van de twaalfde afdeling van de tweede titel van het eerste Boek en in afwijking van artikel 843a, eerste lid, wordt de in het ongelijk gestelde partij desgevorderd veroordeeld in redelijke en evenredige gerechtskosten en andere kosten die de in het...

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.



Closing thoughts & demo

Closing thoughts

How can we trust AI outputs?

- Give the model **enough context**: locations, documents, etc..
- Don't use AI as the only source: **it's not a search engine** (but it can use one)
- Don't share client information or personal data on free tiers
- Monitor what the model does: *tool calls, web search, asking questions,...*

AI is (nog) geen stagiair

Wat taalmodellen wel en niet kunnen

Dr. ing. Pieter Delobelle

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

Slides: pieter.ai/appearances.html