LLaMA's en GEITjes

Successen en uitdagingen voor Nederlandse generatieve taalmodellen

Edwin Rijgersberg – e.rijgersberg@nfi.nl
Nederlands Forensisch Instituut

Closed vs Open

“Kleine” taalmodellen

Google

BERT

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

arXiv:1810.04805v2 [cs.CL] 24 May 2019

Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

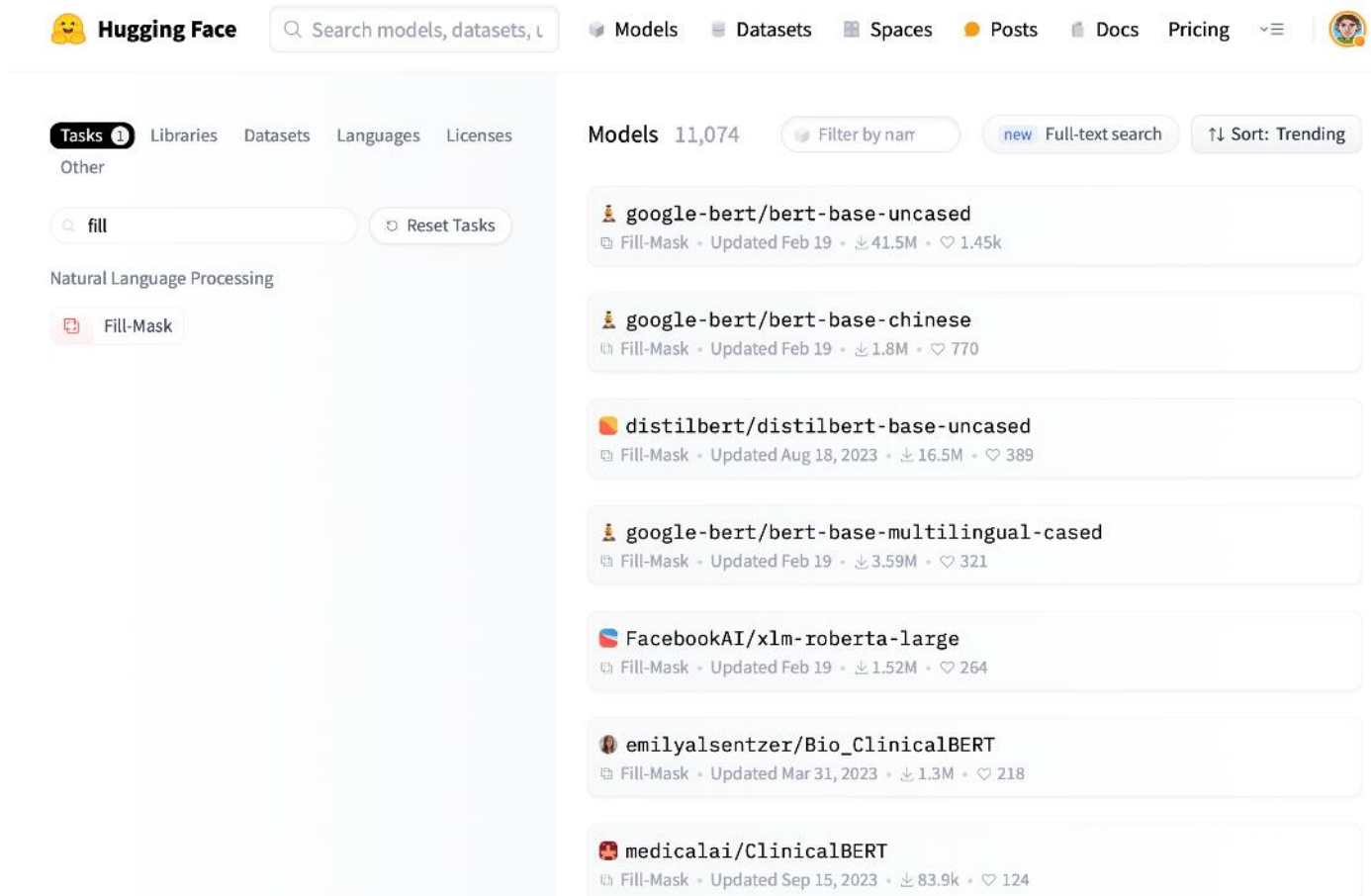
1 Introduction

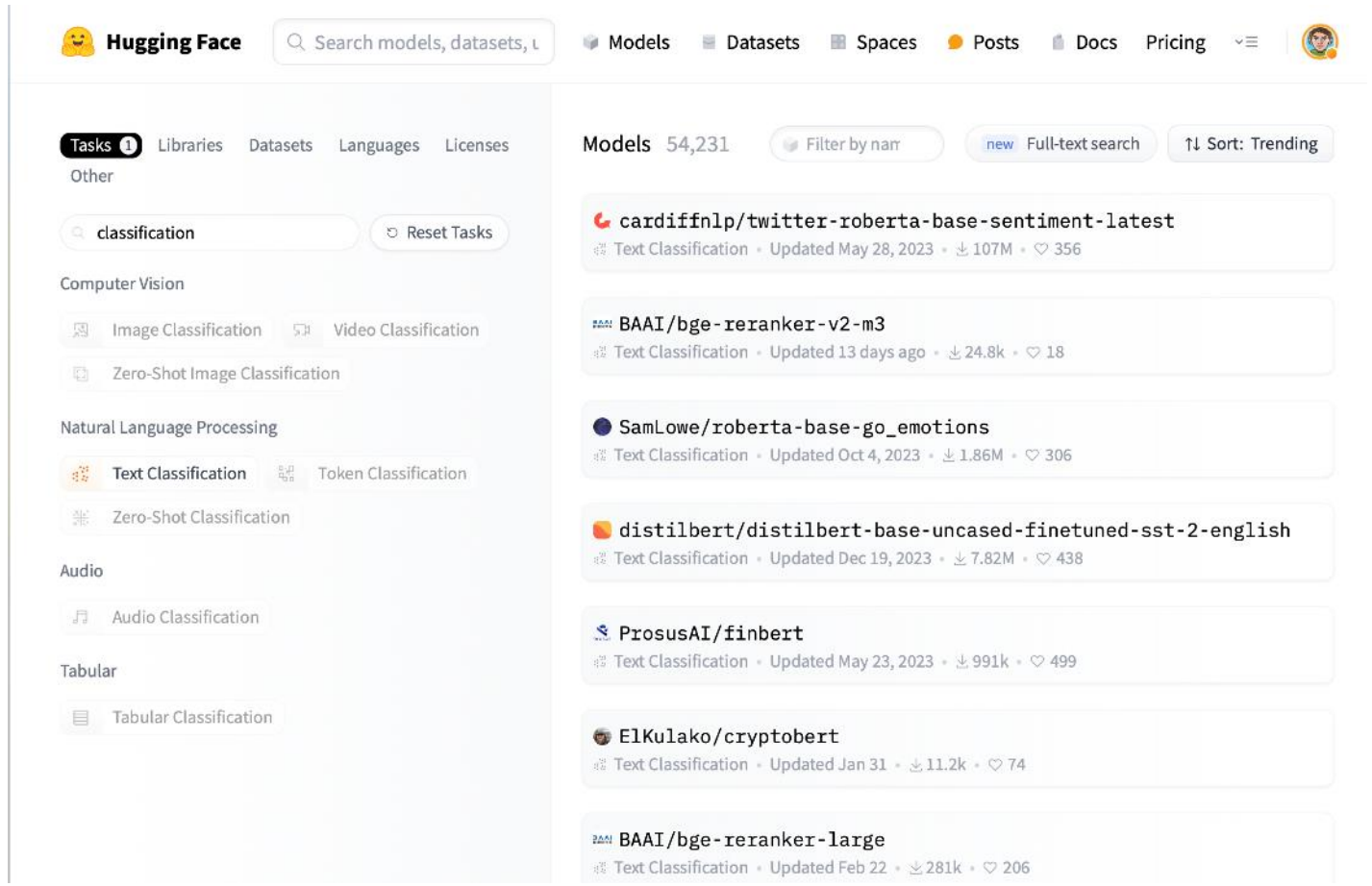
Language model pre-training has been shown to be effective for improving many natural language processing tasks (Dai and Le, 2015; Peters et al., 2018a; Radford et al., 2018; Howard and Ruder, 2018). These include sentence-level tasks such as natural language inference (Bowman et al., 2015; Williams et al., 2018) and paraphrasing (Dolan and Brockett, 2005), which aim to predict the relationships between sentences by analyzing them holistically, as well as token-level tasks such as named entity recognition and question answering, where models are required to produce fine-grained output at the token level (Tjong Kim Sang and De Meulder, 2003; Rajpurkar et al., 2016).

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

We argue that current techniques restrict the power of the pre-trained representations, especially for the fine-tuning approaches. The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in OpenAI GPT, the authors use a left-to-right architecture, where every token can only attend to previous tokens in the self-attention layers of the Transformer (Vaswani et al., 2017). Such restrictions are sub-optimal for sentence-level tasks, and could be very harmful when applying fine-tuning based approaches to token-level tasks such as question answering, where it is crucial to incorporate context from both directions.

In this paper, we improve the fine-tuning based approaches by proposing BERT: **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. BERT alleviates the previously mentioned unidirectionality constraint by using a “masked language model” (MLM) pre-training objective, inspired by the Cloze task (Taylor, 1953). The masked language model randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked


Hugging Face
Search models, datasets, u
Models
Datasets
Spaces
Posts
Docs
Pricing
Tasks 1
Libraries
Datasets
Languages
Licenses
Other
fill
Reset Tasks
Natural Language Processing
Fill-Mask
Models 11,074
Filter by nam
new Full-text search
↑↓ Sort: Trending
google-bert/bert-base-uncased
Fill-Mask • Updated Feb 19 • 41.5M • 1.45k
google-bert/bert-base-chinese
Fill-Mask • Updated Feb 19 • 1.8M • 770
distilbert/distilbert-base-uncased
Fill-Mask • Updated Aug 18, 2023 • 16.5M • 389
google-bert/bert-base-multilingual-cased
Fill-Mask • Updated Feb 19 • 3.59M • 321
FacebookAI/xlm-roberta-large
Fill-Mask • Updated Feb 19 • 1.52M • 264
emilyalsentzer/Bio_ClinicalBERT
Fill-Mask • Updated Mar 31, 2023 • 1.3M • 218
medicalai/ClinicalBERT
Fill-Mask • Updated Sep 15, 2023 • 83.9k • 124

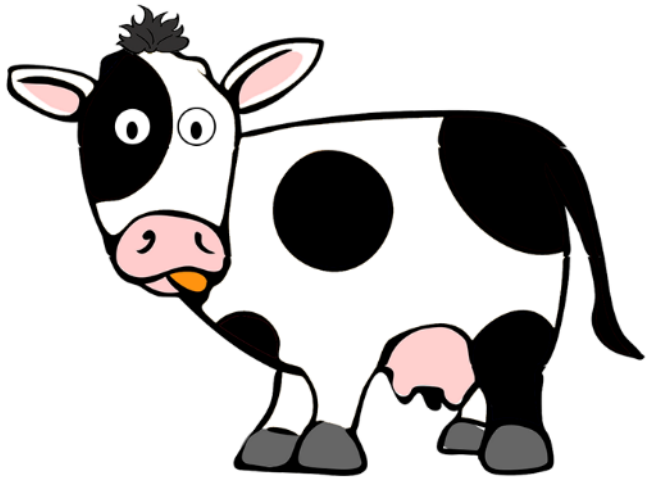

Hugging Face
Search models, datasets, u
Models
Datasets
Spaces
Posts
Docs
Pricing
Tasks 1
Libraries
Datasets
Languages
Licenses
Other
classification
Reset Tasks
Computer Vision
Image Classification
Video Classification
Zero-Shot Image Classification
Natural Language Processing
Text Classification
Token Classification
Zero-Shot Classification
Audio
Audio Classification
Tabular
Tabular Classification
Models 54,231
Filter by nam
new Full-text search
Sort: Trending
cardiffnlp/twitter-roberta-base-sentiment-latest
Text Classification • Updated May 28, 2023 • 107M • 356
BAAI/bge-reranker-v2-m3
Text Classification • Updated 13 days ago • 24.8k • 18
SamLowe/roberta-base-go_emotions
Text Classification • Updated Oct 4, 2023 • 1.86M • 306
distilbert/distilbert-base-uncased-finetuned-sst-2-english
Text Classification • Updated Dec 19, 2023 • 7.82M • 438
ProsusAI/finbert
Text Classification • Updated May 23, 2023 • 991k • 499
ElKulako/cryptobert
Text Classification • Updated Jan 31 • 11.2k • 74
BAAI/bge-reranker-large
Text Classification • Updated Feb 22 • 281k • 206

BERTje: A Dutch BERT model

Wietse de Vries • Andreas van Cranenburgh • Arianna Bisazza • Tommaso Caselli • Gertjan van Noord • Malvina Nissim

Model description

BERTje is a Dutch pre-trained BERT model developed at the University of Groningen.

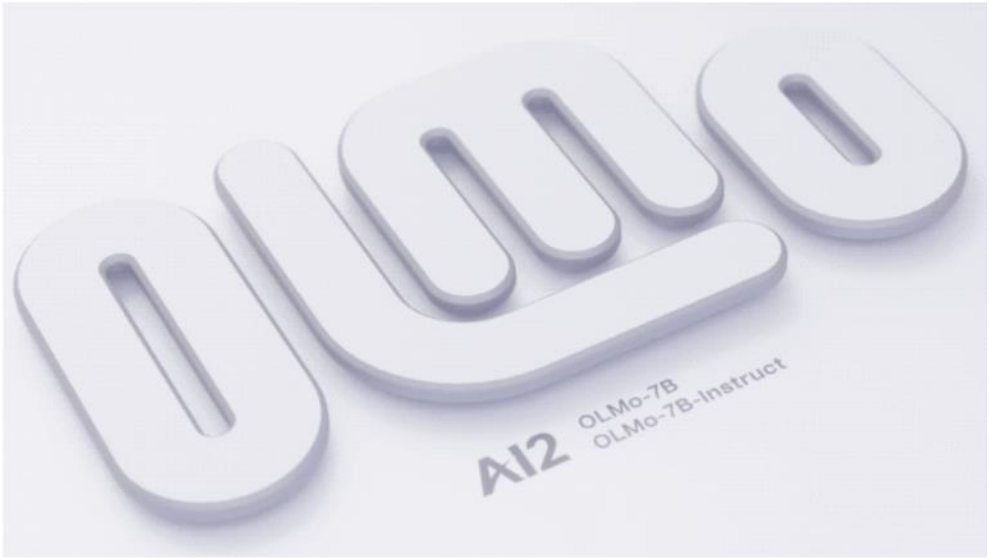


RobBERT

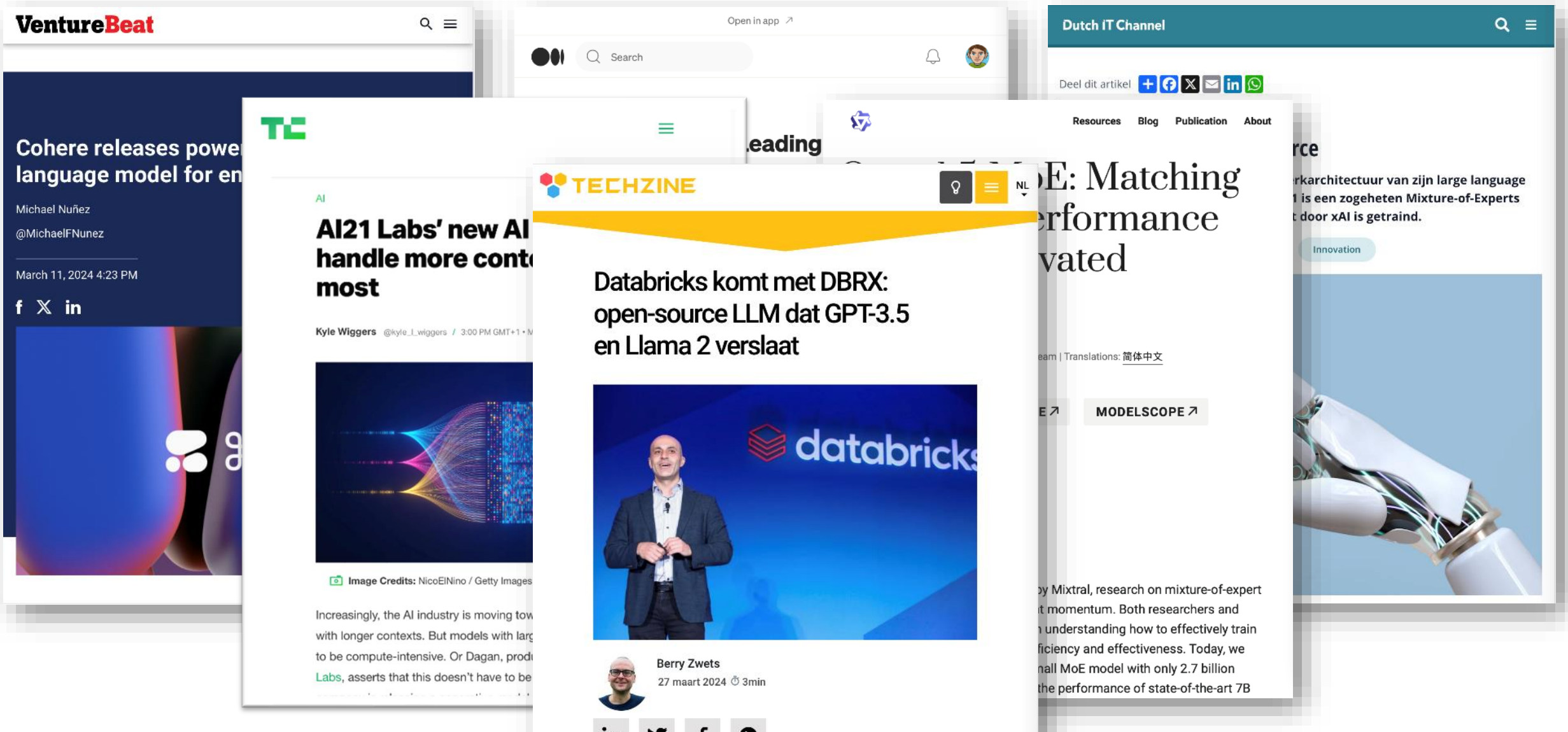
A Dutch RoBERTa-based Language Model

ChatGPT

Reactie open source modellen

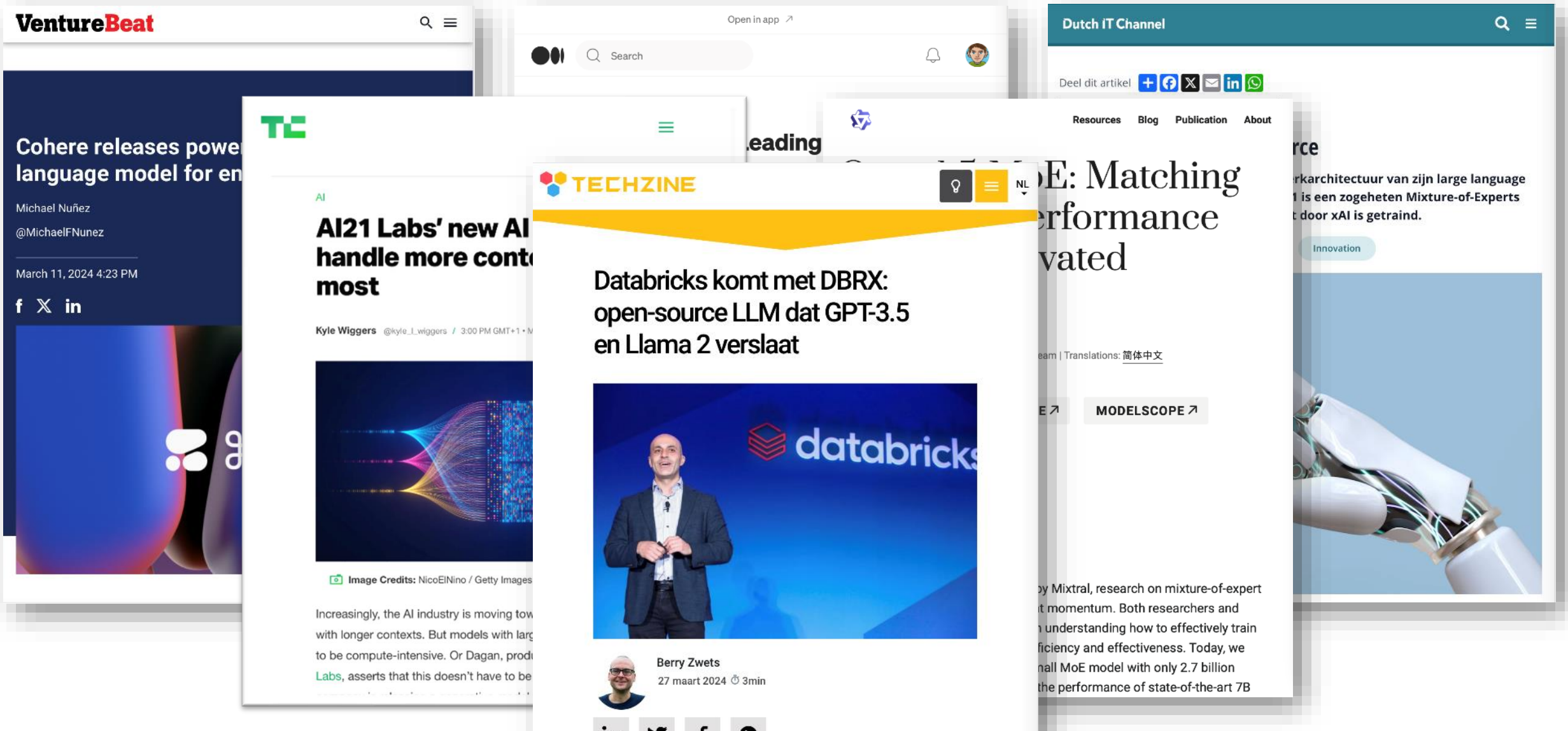


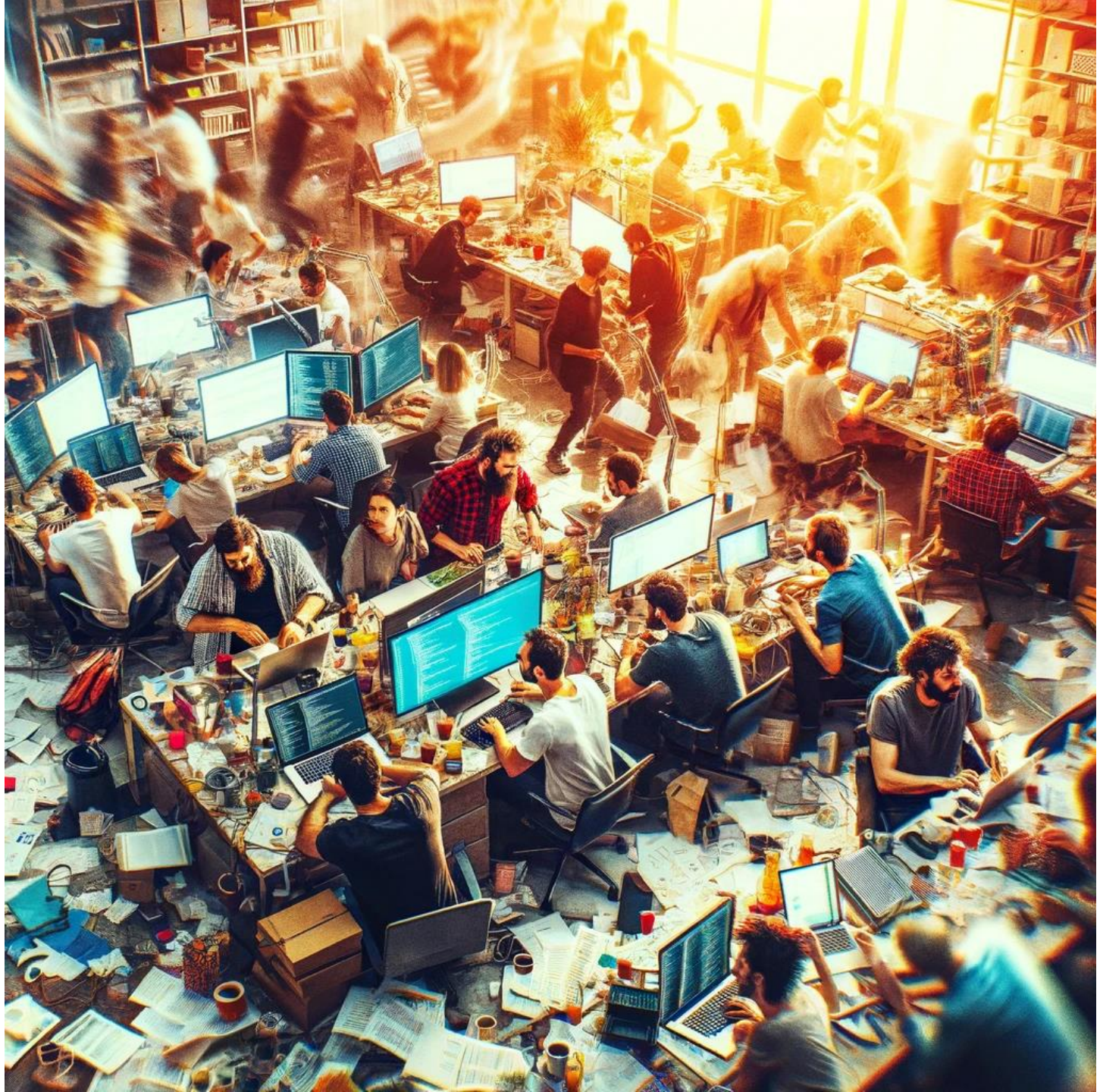
Afgelopen jaar



Afgelopen ~~jaar~~ maand

https://twitter.com/osanseviero/status/1773828987866714570





Waarom open modellen?

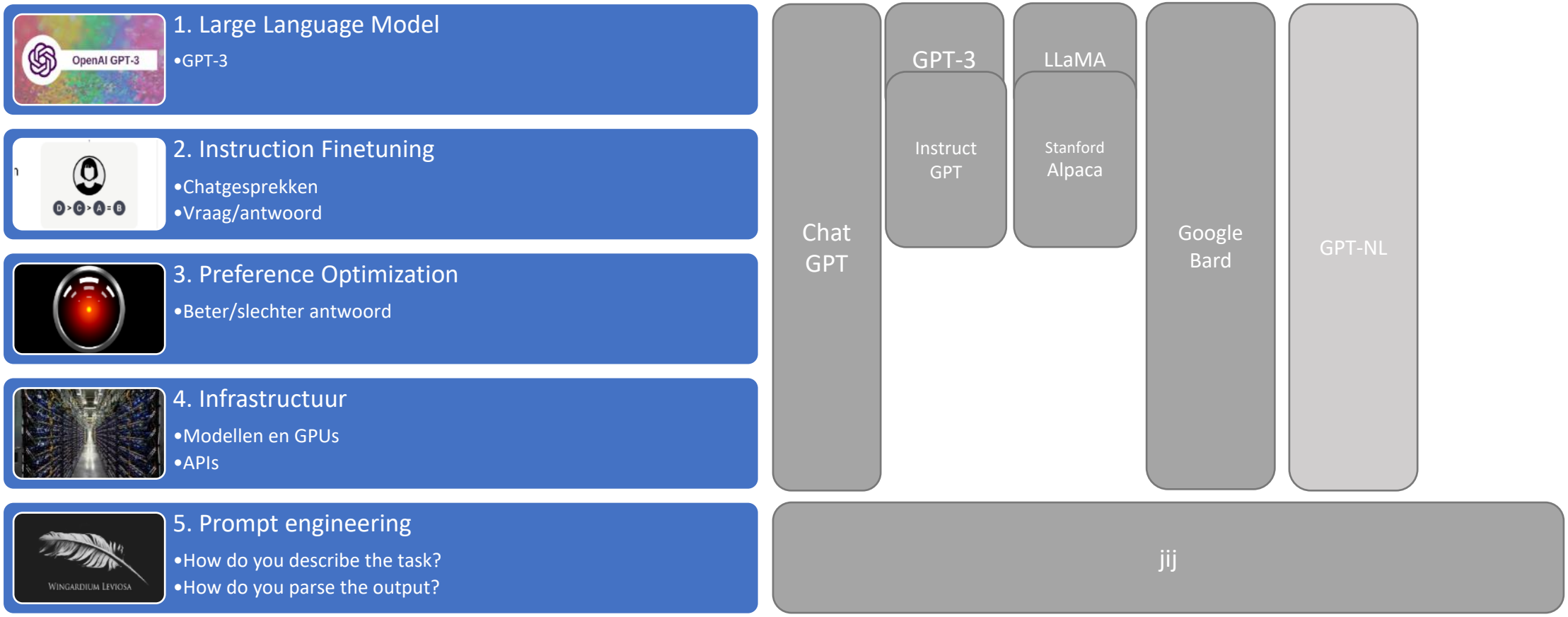
Voordelen

- Eigen infra
  - Data vertrouwelijk
- Onderzoeken / valideren
- Meer keuze
- Finetunen voor specifieke taken
- Bouw voort op elkaars werk

Nadelen

- Eigen infra
- Minder “intelligent”
- Minder overzicht
- Licentie kan problematisch zijn

Hoe werkt een taalmodel?



GEITje 7B

een groot open Nederlands taalmodel

Wat is GEITje?

README | Apache-2.0 license

🐐 GEITje 7B: een groot open Nederlands taalmodel

📄 English README | 🤖 GEITje-chat-v2 demo

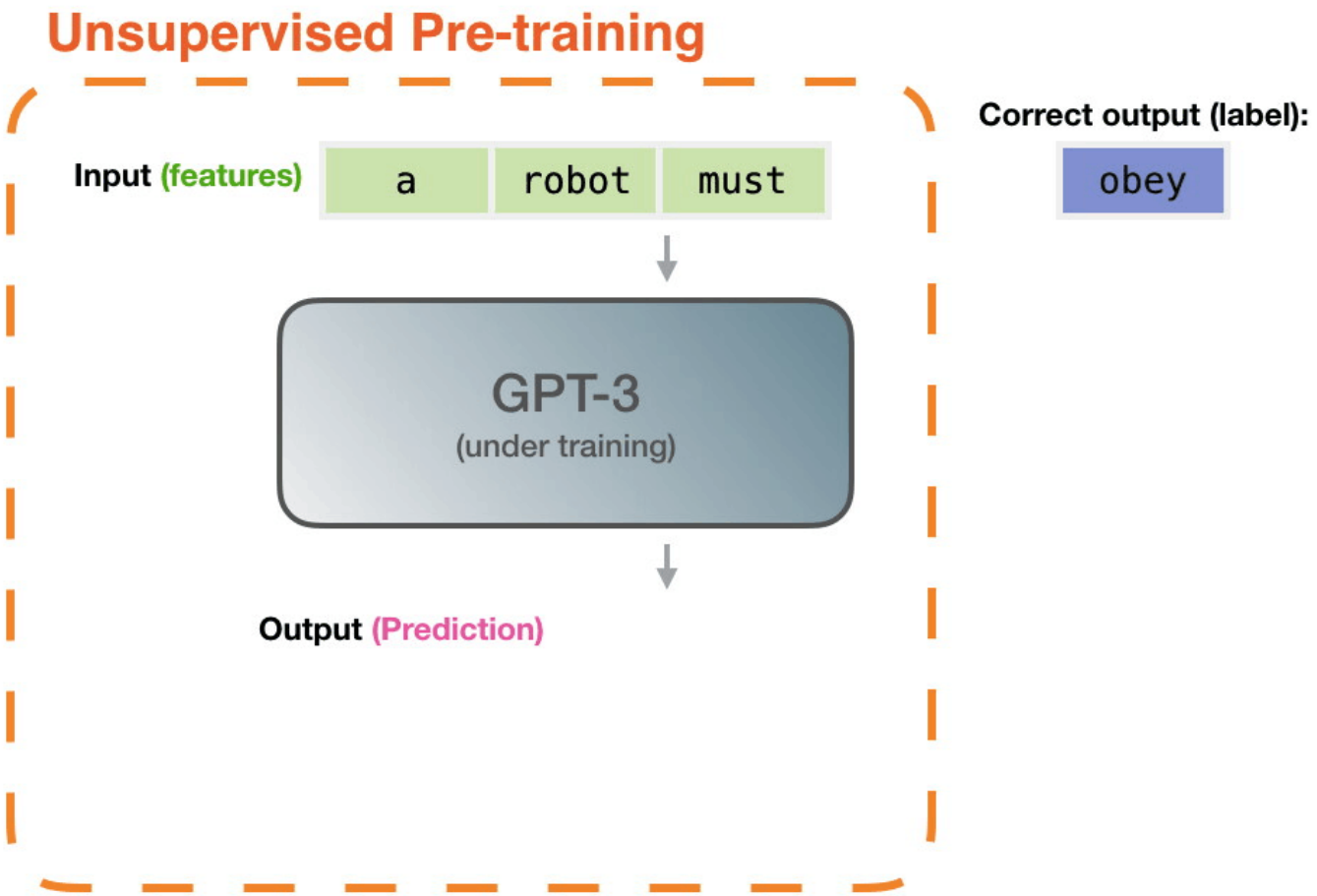
GEITje is een Nederlandstalig groot open taalmodel met 7 miljard parameters, gebaseerd op Mistral 7B. Het is (verder) getraind op 10 miljard tokens aan Nederlandstalige tekst. Daardoor heeft het beter Nederlands geleerd, en meer kennis over Nederlandse onderwerpen.

Hoe train je een taalmodel?

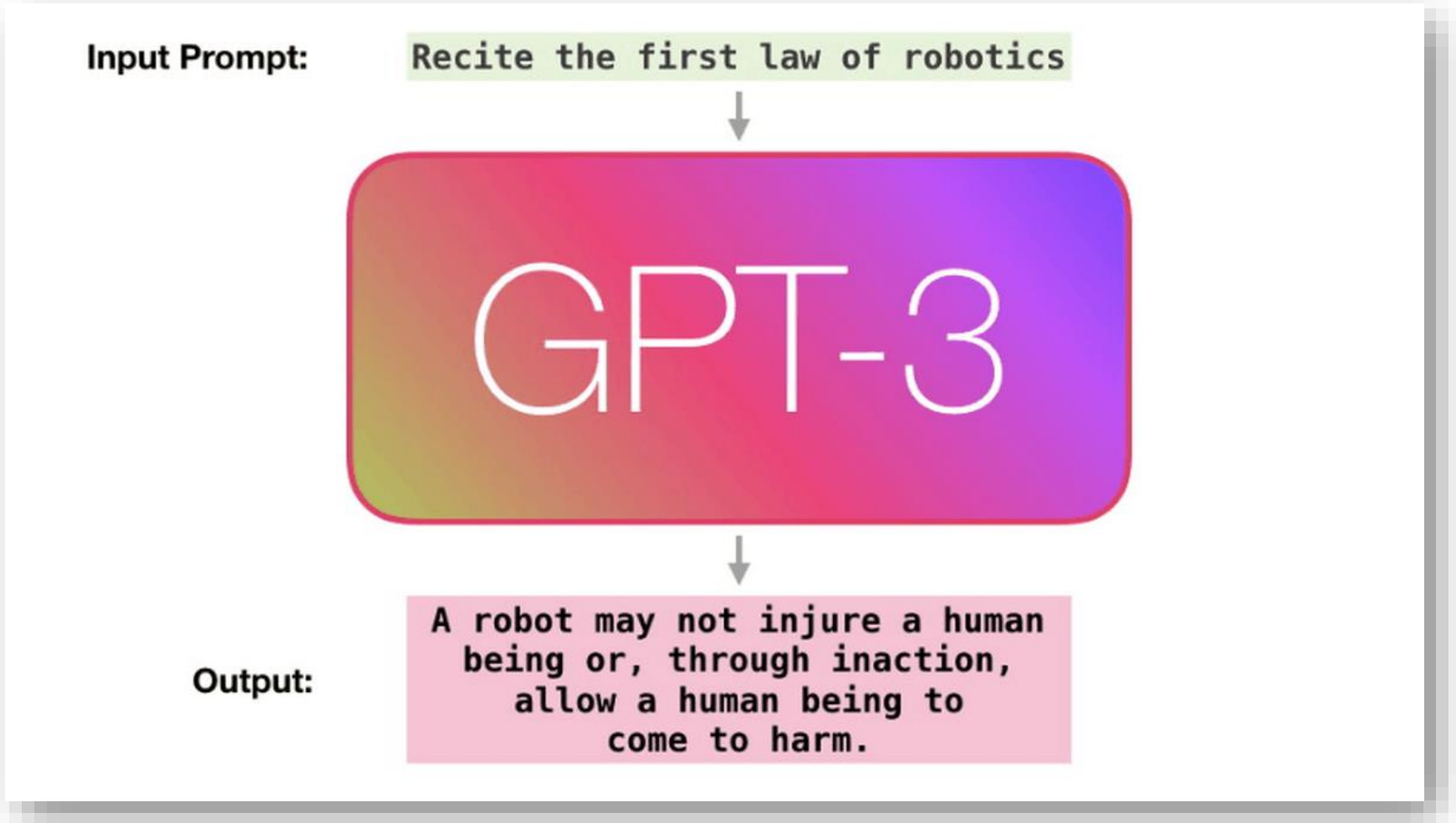
Text: Second Law of Robotics: A robot must obey the orders given it by human beings

Generated training examples

Example #	Input (features)	Correct output (labels)
1	Second law of robotics :	a
2	Second law of robotics : a	robot
3	Second law of robotics : a robot	must
...		



https://gigazine.net/gsc_news/en/20200729-how-gpt-3-work/



https://gigazine.net/gsc_news/en/20200729-how-gpt-3-work/

Mistral 7B

MISTRAL AI_

- 7B model, beter dan Llama 2 7B en 13B
- Open-source licentie: Apache-2.0
- Trainingsdata: compleet onbekend
  - Blijkbaar ook wat Nederlands?

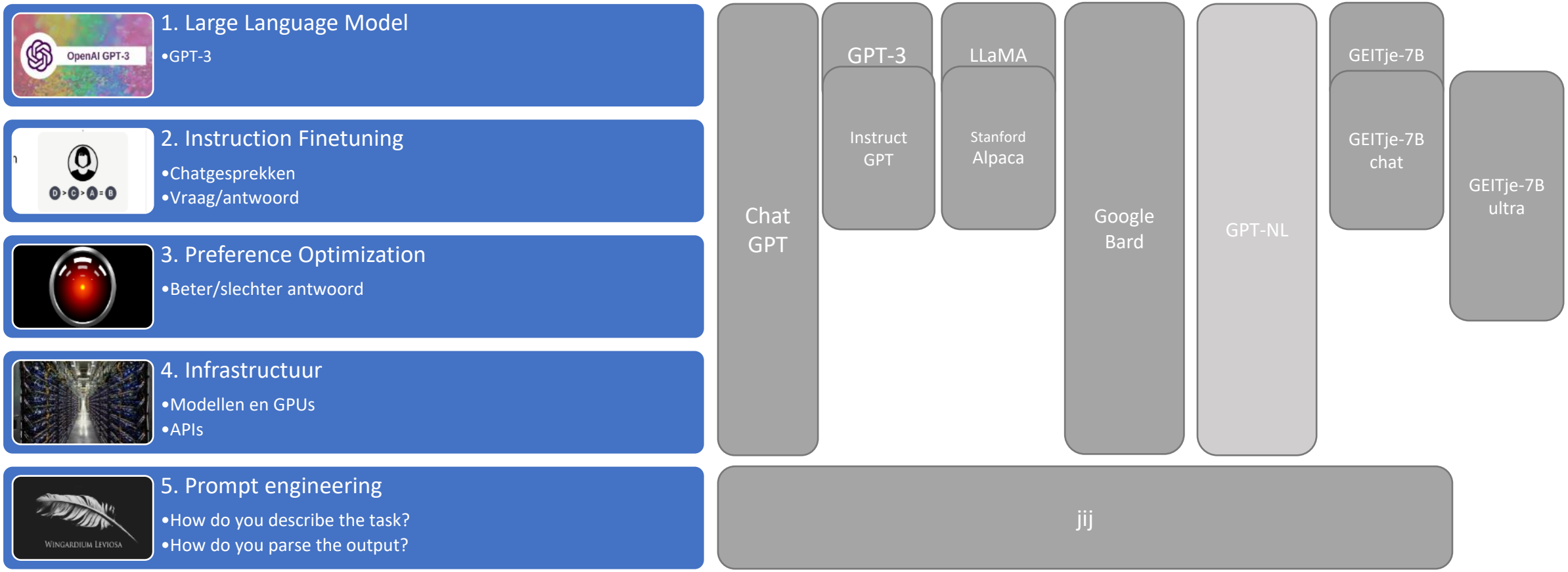
- Chat-model:
  - mistralai/**Mistral-7B-Instruct-v0.1**
  - < honderden andere finetunes >

GEITje 7B

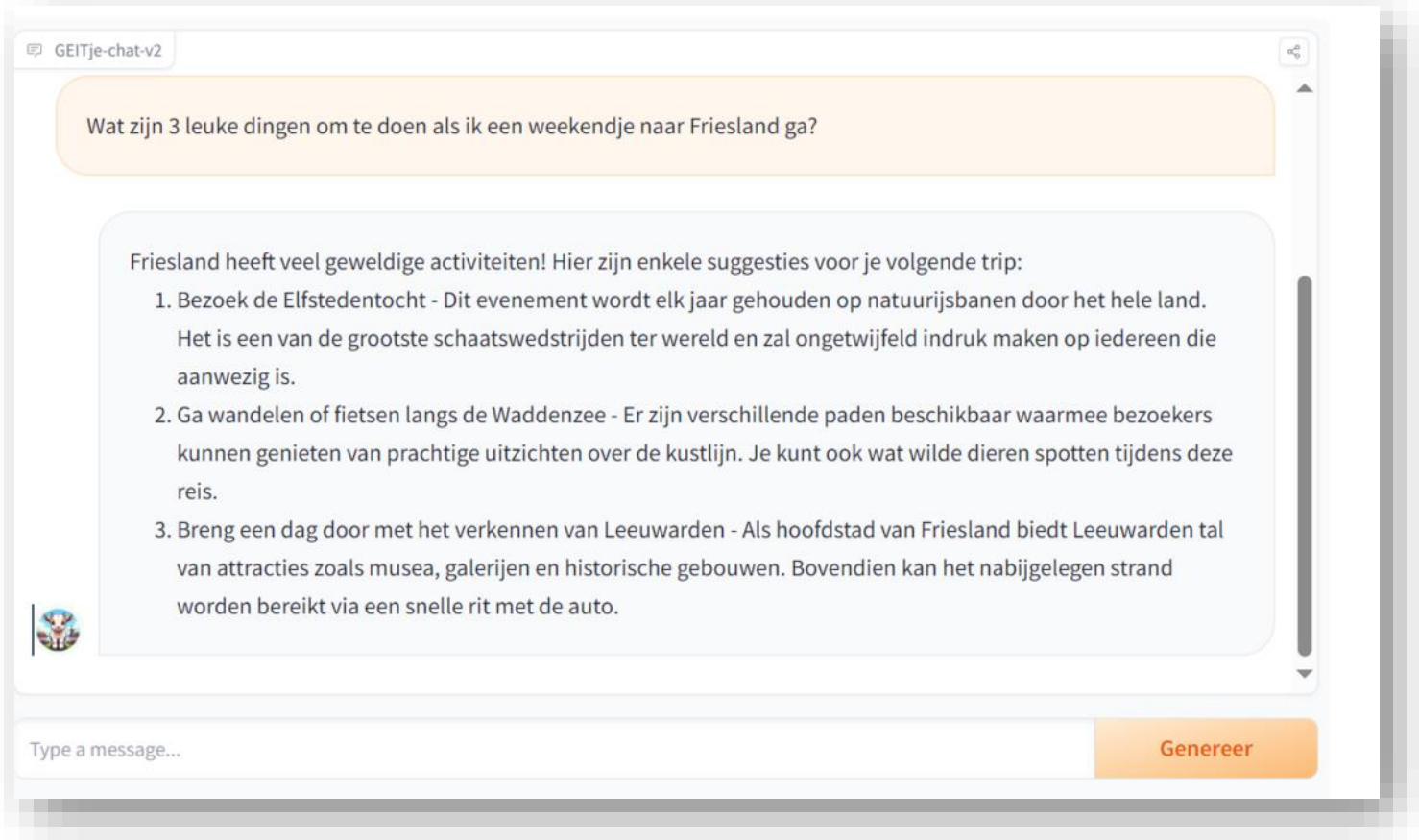
- Doorgetraind op 10 miljard tokens aan NL tekst
  - Gigacorpus NL
  - MADLAD-400
- Open-source licentie : Apache-2.0
- Betere kennis van:
  - Nederlandse taal
  - Nederlandse cultuur

- Chat-modellen:
  - Rijgersberg/**GEITje-7B-chat-v2**
  - BramVanroy/**GEITje-7B-ultra**

Hoe werkt een taalmodel?



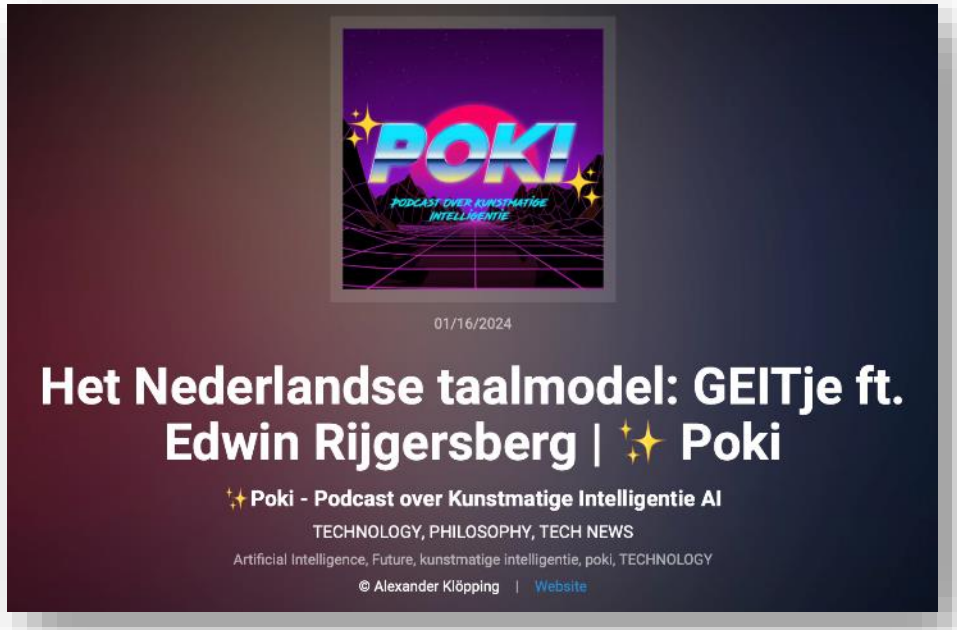
Demo



https://huggingface.co/spaces/BramVanroy/GEITje-7B-ultra

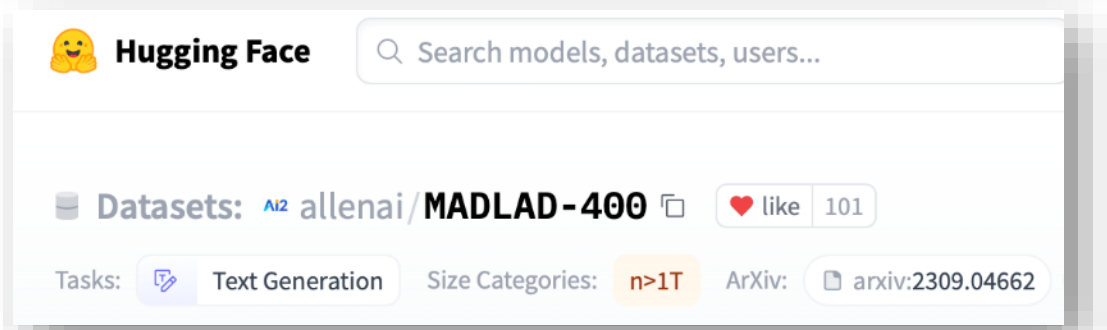
Meer weten?

- De hoofdpagina: github.com/Rijgersberg/GEITje
- Demo's:
  - huggingface.co/spaces/Rijgersberg/GEITje-7B-chat
  - huggingface.co/spaces/BramVanroy/GEITje-7B-ultra
- Mijn blog: goingdutch.ai
- POKI Podcast: art19.com/shows/poki/episodes/0d22d575-d581-480d-80fb-5848bcb9e5d6



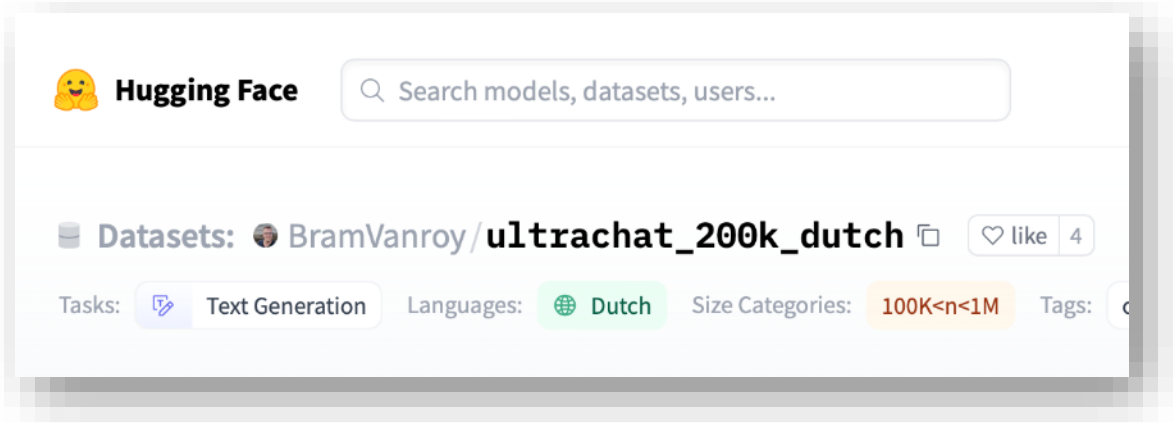
Data – Pretrainen

- Het Nederlandse Gigacorpus *(Bob Lucassen)*
  - Fora
  - Rechtspraak
  - Twitter
  - Nieuwsartikelen
  - …
- MADLAD-400 *(Google)*
  - Gefilterd uit CommonCrawl
  - NL-subset



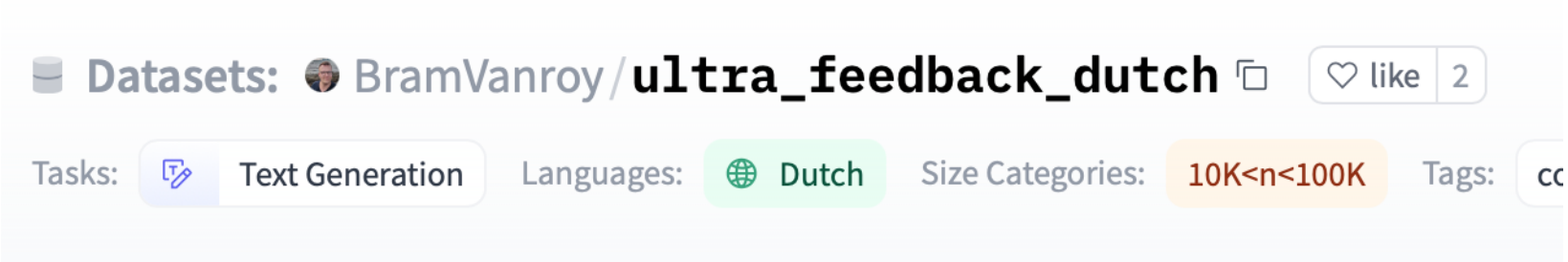
Data – Finetuning

- Er bestaat geen Nederlandse data
- Synthetische data:
  1. Vertaal alle prompts van UltraChat200k
  2. Stel persona's op voor verschilllende types gebruikers
  3. Laat GPT-4 conversaties schrijven
- GPT-NL gaat echte data maken



Data – Preference optimazation

- Er bestaat geen Nederlandse data
- Synthetische data:
  1. Vertaal alle prompts van UltraFeedback
  2. Genereer een antwoord met GEITje-chat: **fout**
  3. Genereer een antwoord met GPT-4: **goed**
- GPT-NL gaat echte data maken



Gebruik

- Taalwetenschappers: experimenten
- Betere chatbot (DPO-training)
- Opdrachten in de stijl van scholieren.nl
- Extraheren van diagnoses in medische documenten
- Verkennen van forensische toepassingen
- …?

Volgende stappen – Occiglot-7b-nl-en

Occiglot

Home » Posts

Announcing Occiglot: Polyglot Language Models for the Occident

Today, we announce Occiglot: A large-scale research collective for open-source development of Large Language Models by and for Europe.

March 6, 2024 · Occiglot Team

occiglot / **occiglot-7b-de-en** like 4

Text Generation | Transformers | Safetensors | English | German | mistral

Collaborators

dfki ai | hessian.AI | Technische Universität Darmstadt

Common Crawl | pleias | EleutherAI

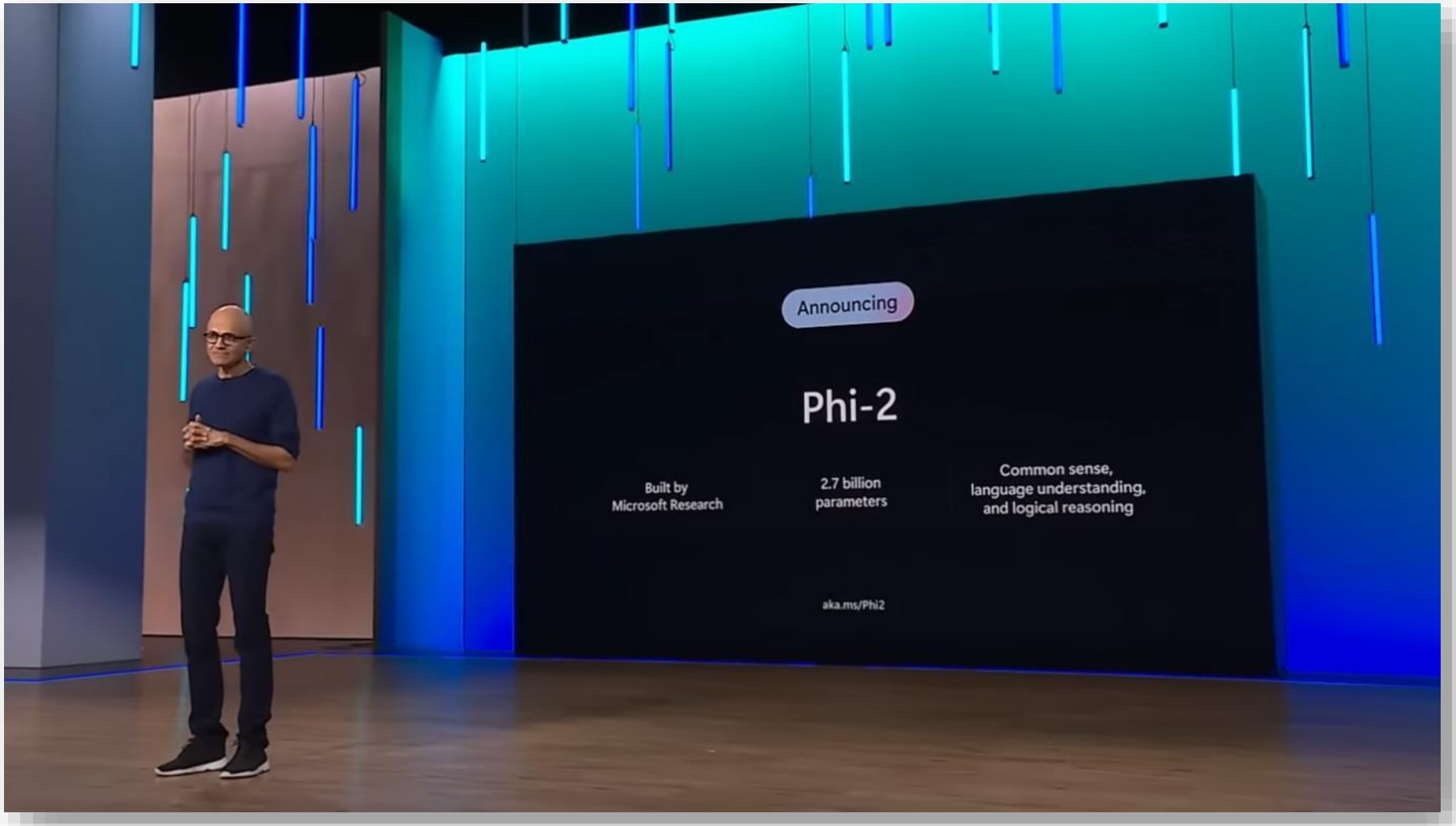
DISCO Research | BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación | European Language Grid | European Language Equality

KU Leuven

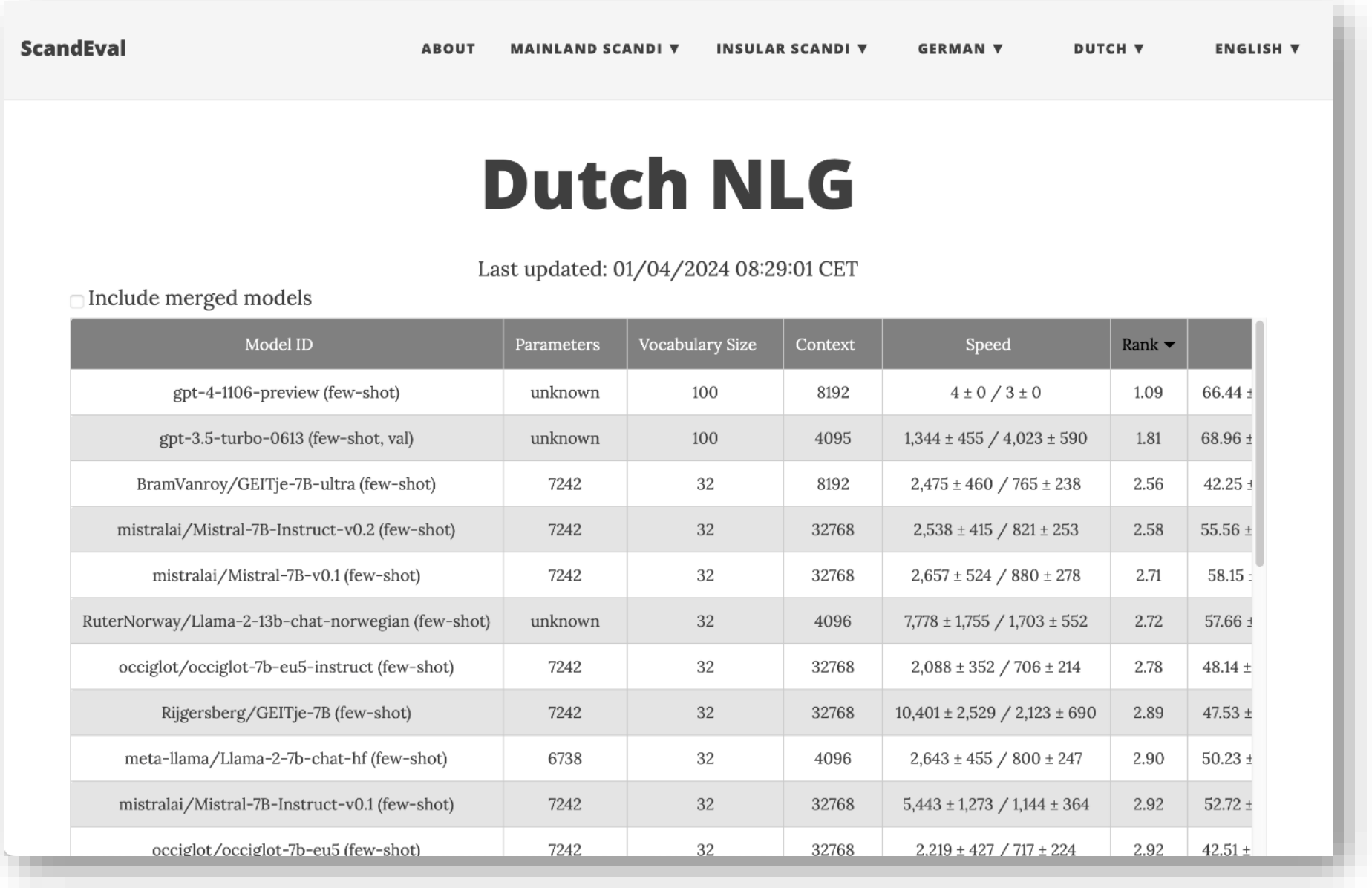
More to be announced soon...

Grotere modellen?

Kleinere modellen?



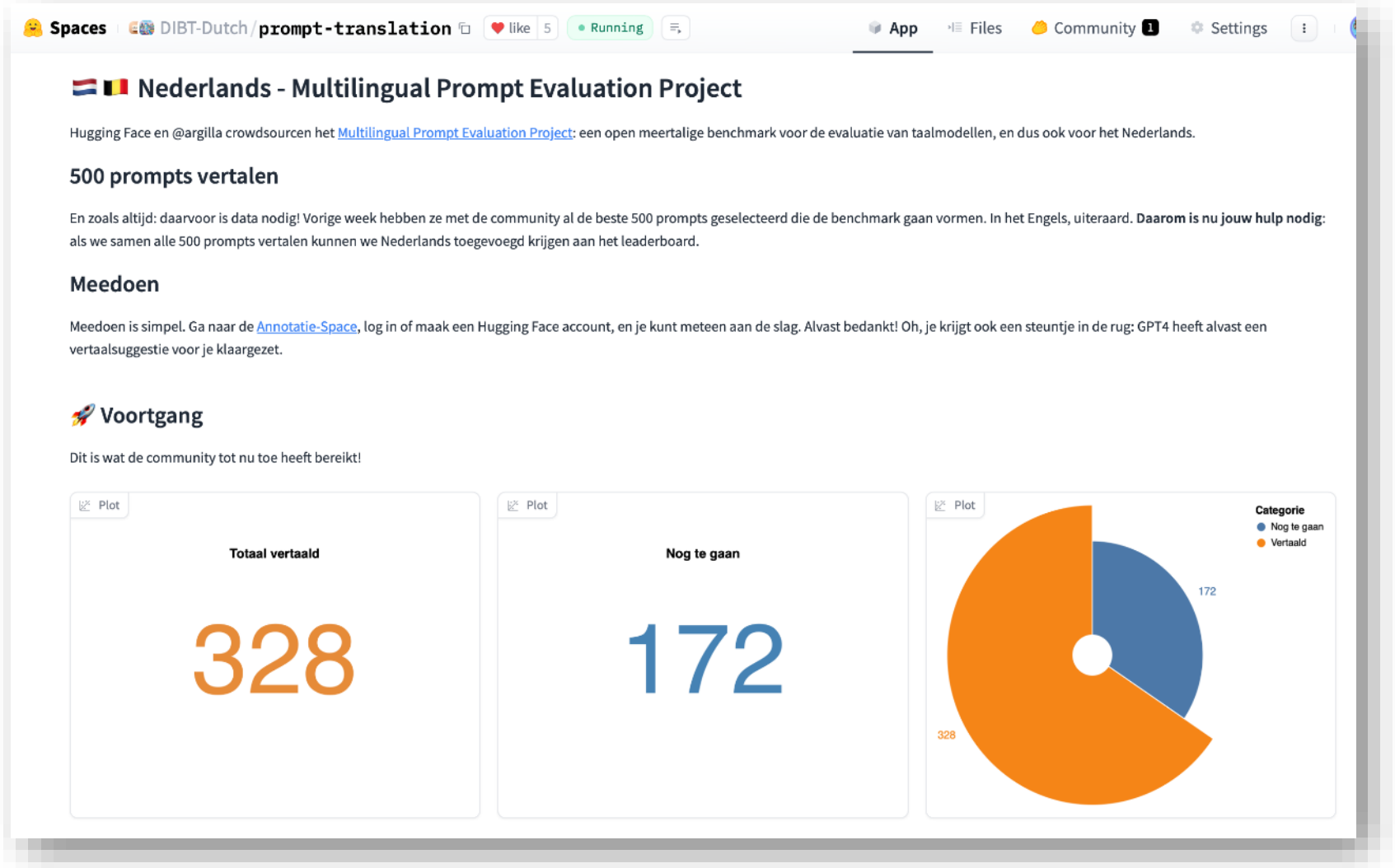
Evaluatie



Model ID	Parameters	Vocabulary Size	Context	Speed	Rank ▼	
gpt-4-1106-preview (few-shot)	unknown	100	8192	4 ± 0 / 3 ± 0	1.09	66.44 ±
gpt-3.5-turbo-0613 (few-shot, val)	unknown	100	4095	1,344 ± 455 / 4,023 ± 590	1.81	68.96 ±
BramVanroy/GEITje-7B-ultra (few-shot)	7242	32	8192	2,475 ± 460 / 765 ± 238	2.56	42.25 ±
mistralai/Mistral-7B-Instruct-v0.2 (few-shot)	7242	32	32768	2,538 ± 415 / 821 ± 253	2.58	55.56 ±
mistralai/Mistral-7B-v0.1 (few-shot)	7242	32	32768	2,657 ± 524 / 880 ± 278	2.71	58.15 ±
RuterNorway/Llama-2-13b-chat-norwegian (few-shot)	unknown	32	4096	7,778 ± 1,755 / 1,703 ± 552	2.72	57.66 ±
occiglot/occiglot-7b-eu5-instruct (few-shot)	7242	32	32768	2,088 ± 352 / 706 ± 214	2.78	48.14 ±
Rijgersberg/GEITje-7B (few-shot)	7242	32	32768	10,401 ± 2,529 / 2,123 ± 690	2.89	47.53 ±
meta-llama/Llama-2-7b-chat-hf (few-shot)	6738	32	4096	2,643 ± 455 / 800 ± 247	2.90	50.23 ±
mistralai/Mistral-7B-Instruct-v0.1 (few-shot)	7242	32	32768	5,443 ± 1,273 / 1,144 ± 364	2.92	52.72 ±
occiglot/occiglot-7b-eu5 (few-shot)	7242	32	32768	2,219 ± 427 / 717 ± 224	2.92	42.51 ±

https://scandeval.com/dutch-nlg/

Multilingual Prompt Evaluation Project



https://huggingface.co/spaces/DIBT-Dutch/prompt-translation

LLaMA's en GEITjes

Successen en uitdagingen voor Nederlandse generatieve taalmodellen

Edwin Rijgersberg – e.rijgersberg@nfi.nl
Nederlands Forensisch Instituut