GPT-NL

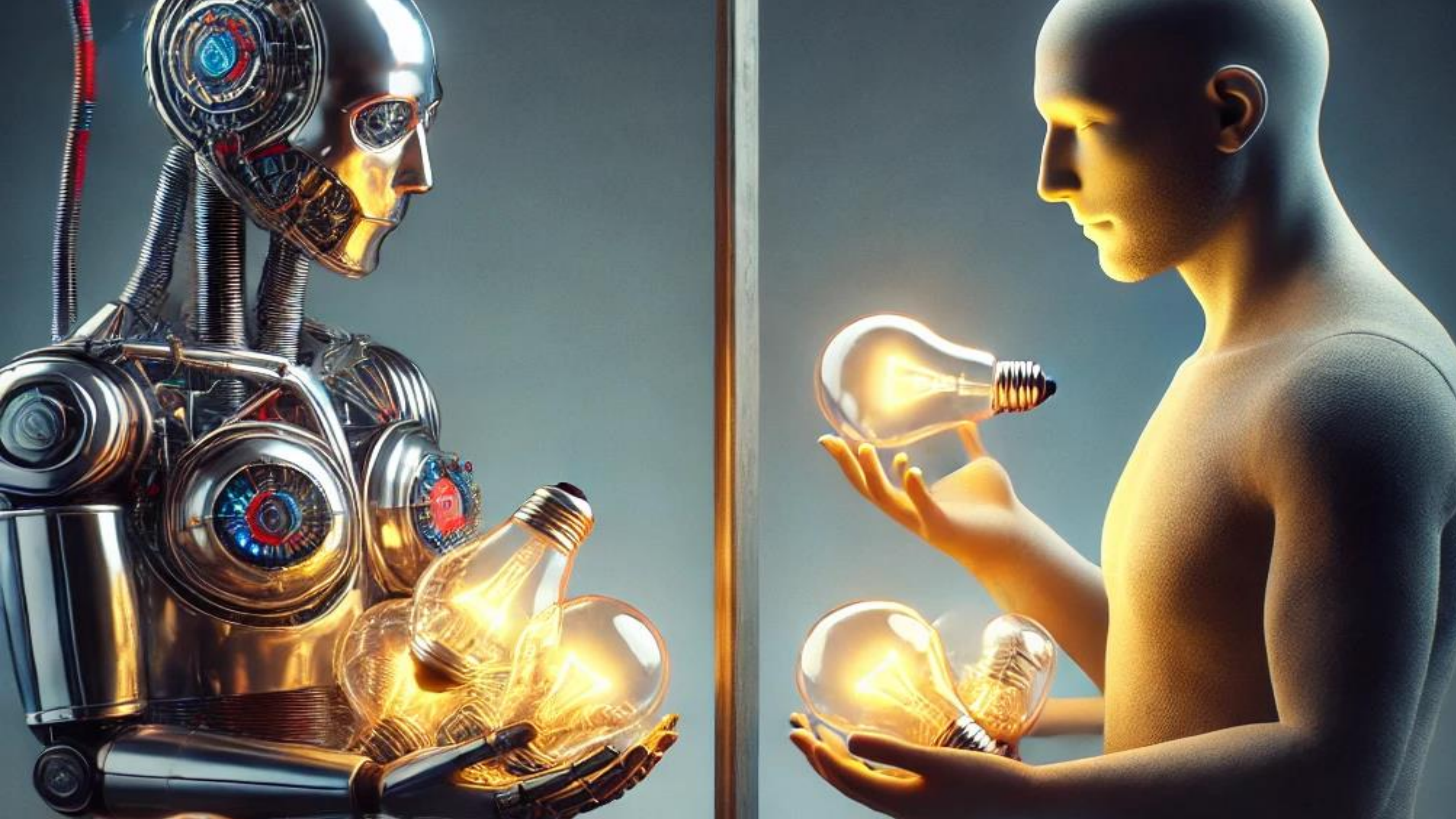
FACILITEIT VOOR EEN SOEVEREIN NEDERLANDS TAALMODEL

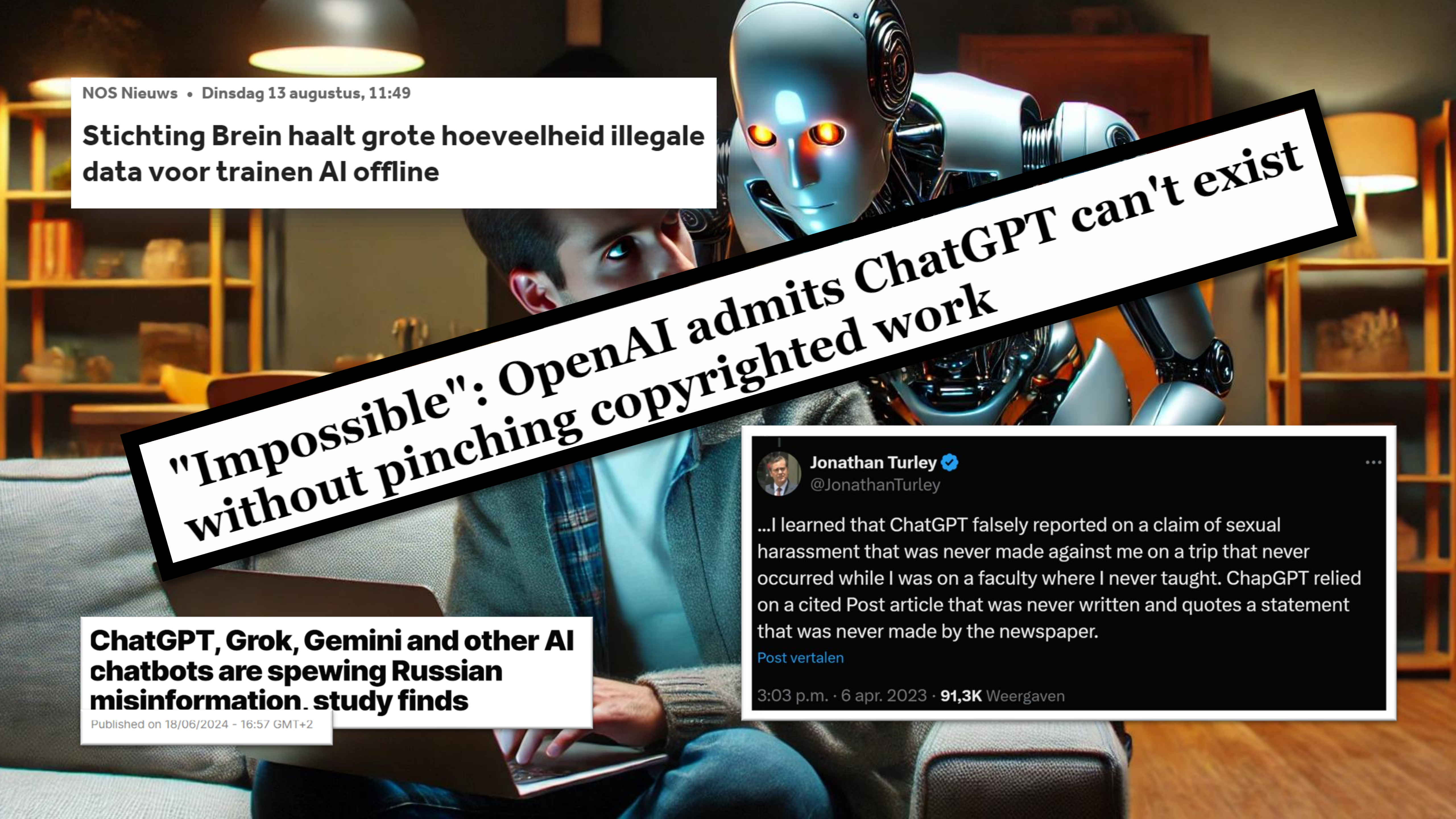
Data Expo 2024

TNO innovation for life

Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

SURF





NOS Nieuws • Dinsdag 13 augustus, 11:49

Stichting Brein haalt grote hoeveelheid illegale data voor trainen AI offline

"Impossible": OpenAI admits ChatGPT can't exist without pinching copyrighted work

ChatGPT, Grok, Gemini and other AI chatbots are spewing Russian misinformation, study finds

Published on 18/06/2024 - 16:57 GMT+2

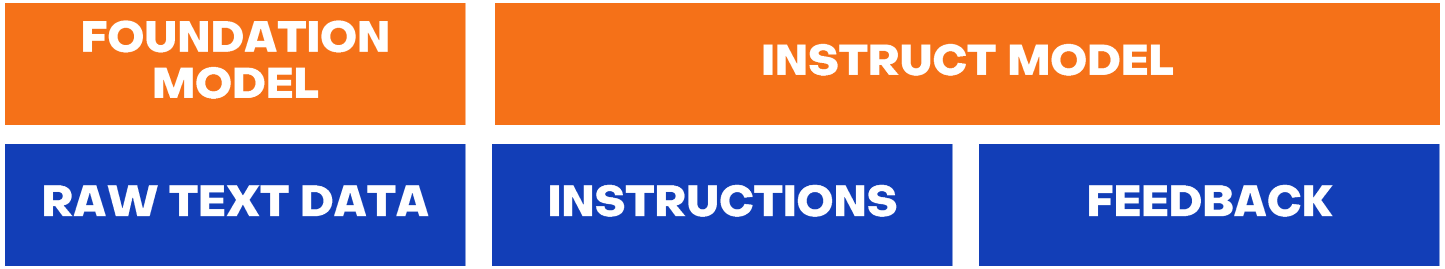
Jonathan Turley
@JonathanTurley

...I learned that ChatGPT falsely reported on a claim of sexual harassment that was never made against me on a trip that never occurred while I was on a faculty where I never taught. ChapGPT relied on a cited Post article that was never written and quotes a statement that was never made by the newspaper.

Post vertalen

3:03 p.m. · 6 apr. 2023 · **91,3K** Weergaven

A **lawful** Dutch-English Large Language Model,
Trained on a dataset we are collecting from scratch,
Using data that we are allowed to use,
Striving to be as transparent and compliant as possible


FOUNDATION MODEL

INSTRUCT MODEL

RAW TEXT DATA

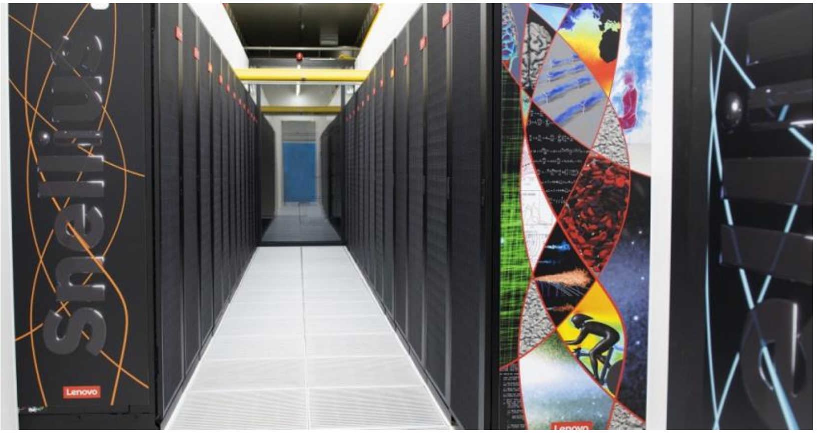
INSTRUCTIONS

FEEDBACK

The GPT-NL consortium

TNO innovation for life

SURF



Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

For whom?

Research institutes

Education

Focus on three main capabilities:

1. Summarisation
2. Simplification
3. Retrieval-Augmented Generation (RAG)

Insurance & banking

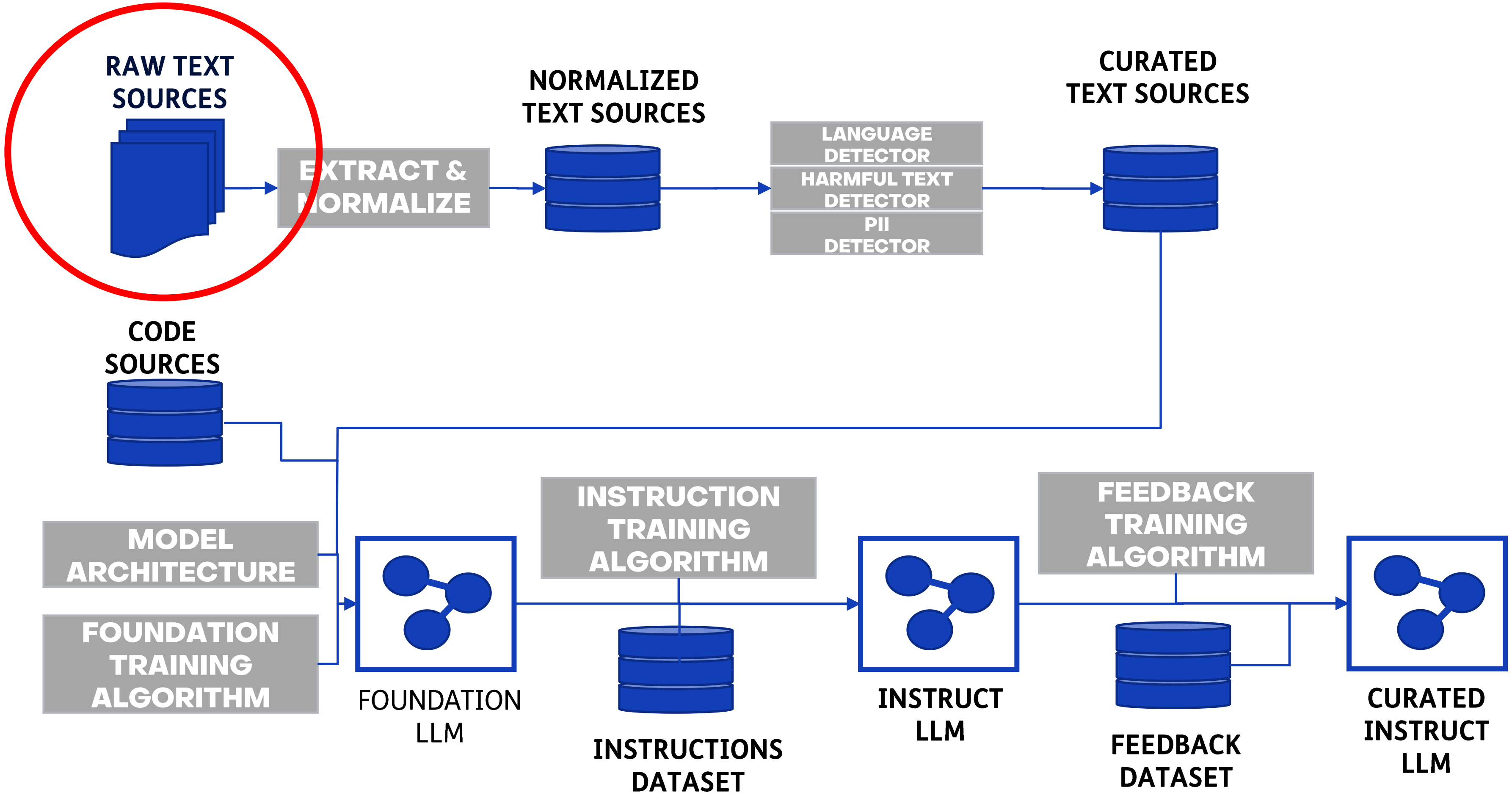
Law enforcement

Defense

Social welfare

GPT-NL

Training pipeline



SURF TNO innovation for life Nederlands Forensisch Instituut Ministerie van Justitie en Veiligheid

Training data

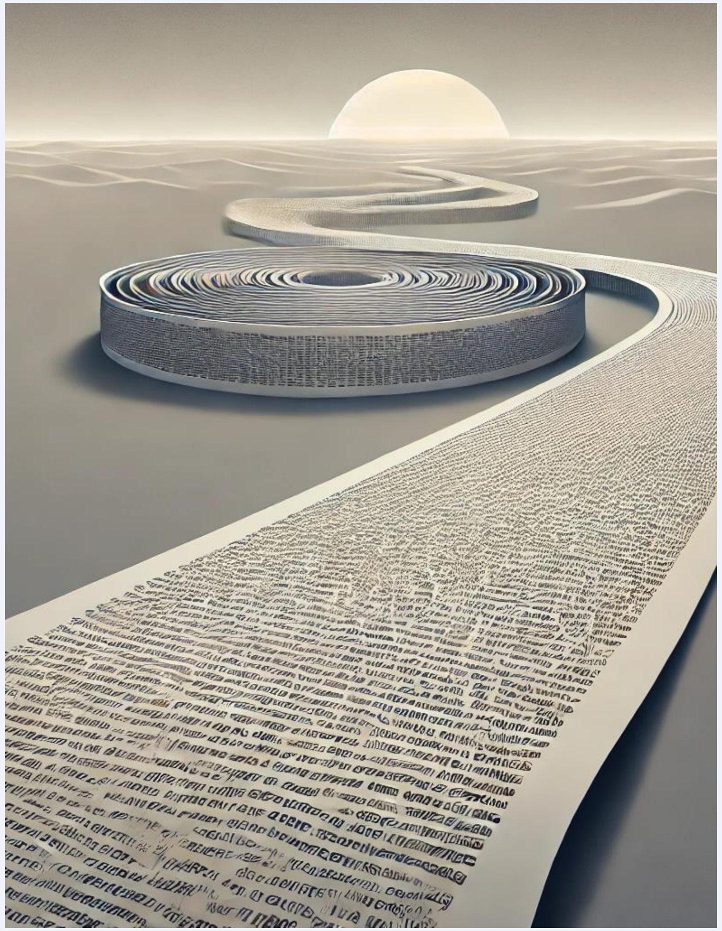
At least 300B tokens …

… ± 3 million x the first Harry Potter book

… ± 6 x all Dutch newspapers and magazines

… ± 5 km of pages when printed

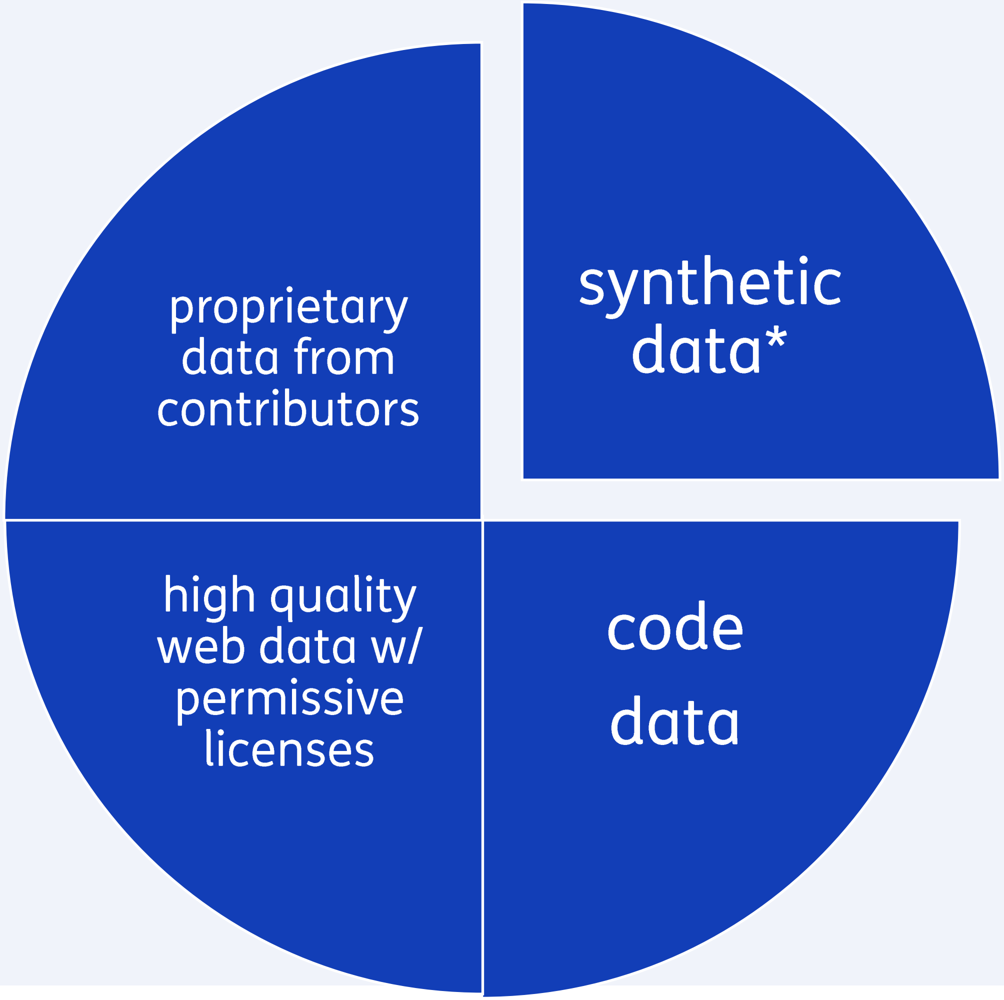
… 2% of Llama 3’s training data



Dall-E imagines large amounts of text

High quality data outweighs more data

* Still under consideration



proprietary data from contributors

synthetic data*

high quality web data w/ permissive licenses

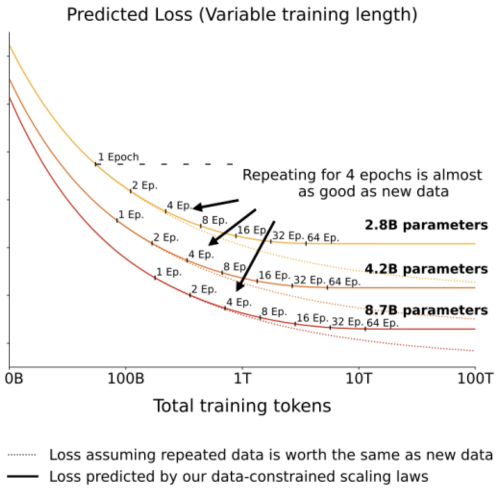
code data

*Tan & Wang, 1.5-Pints Technical Report: Pretraining in Days, Not Months (2024), Gunasekar et al., Textbooks Are All You Need (2023), Sachdeva et al., How to Train Data-Efficient LLMs (2024)

Do more with less

oversampling

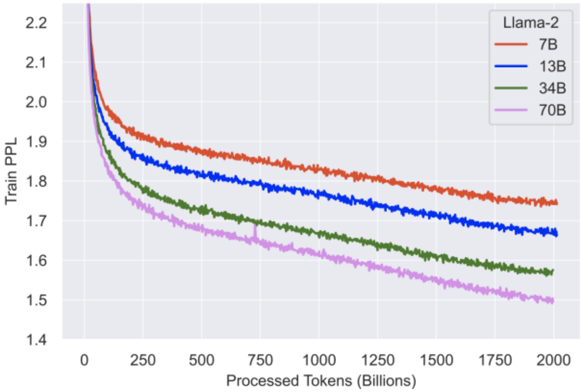
- Multiple training epochs on the same data



Muennighoff et al., Scaling Data-Constrained Language Models (2023)

larger model size

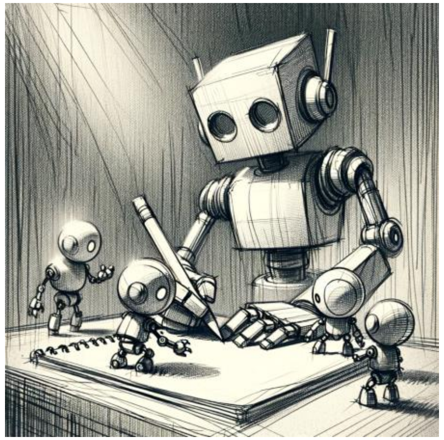
- Larger models are smarter with same number of processed tokens
- However, costlier for inference



Touvron et al, Llama 2: Open Foundation and Fine-Tuned Chat Models (2023)

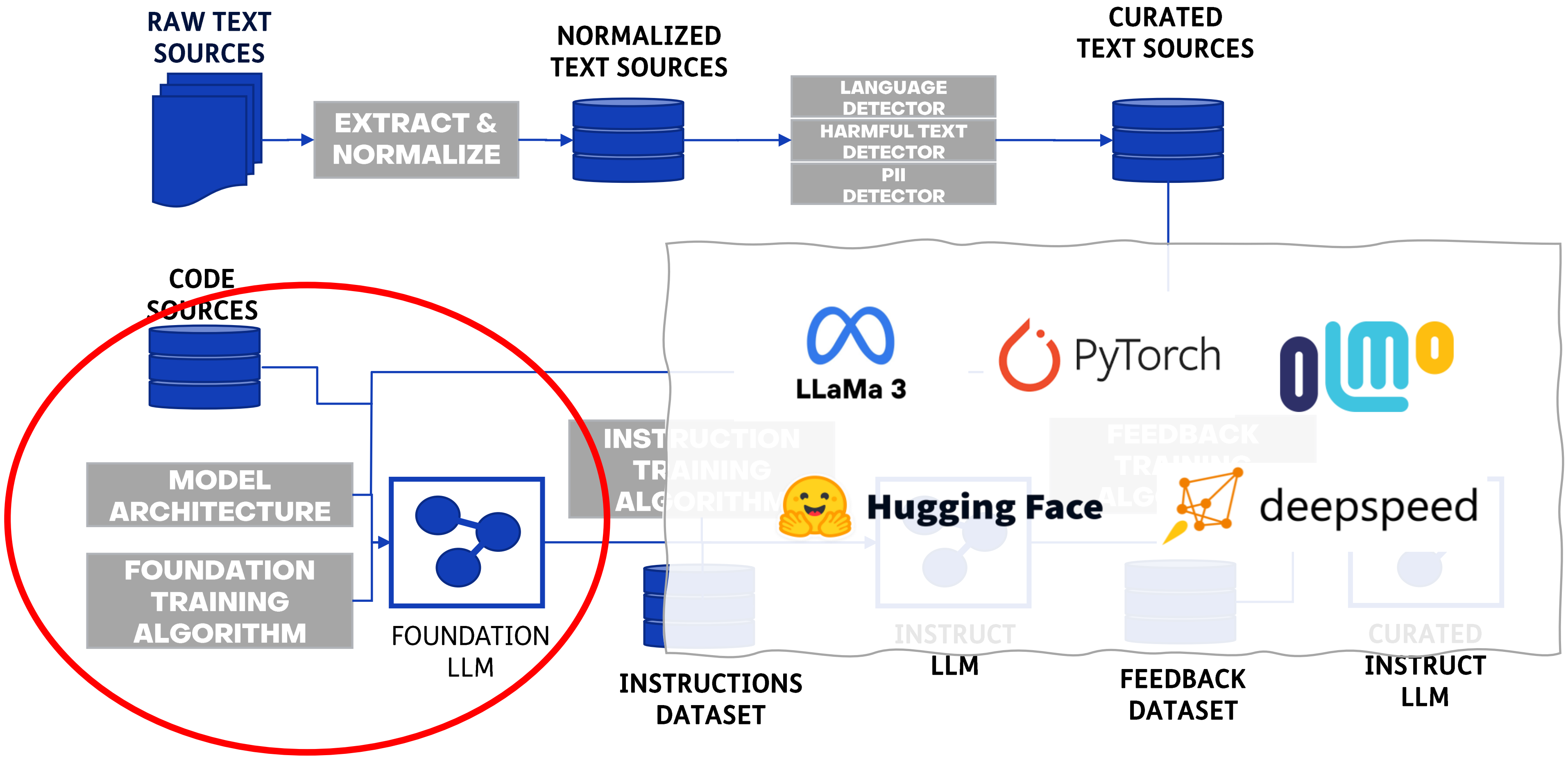
synthesis

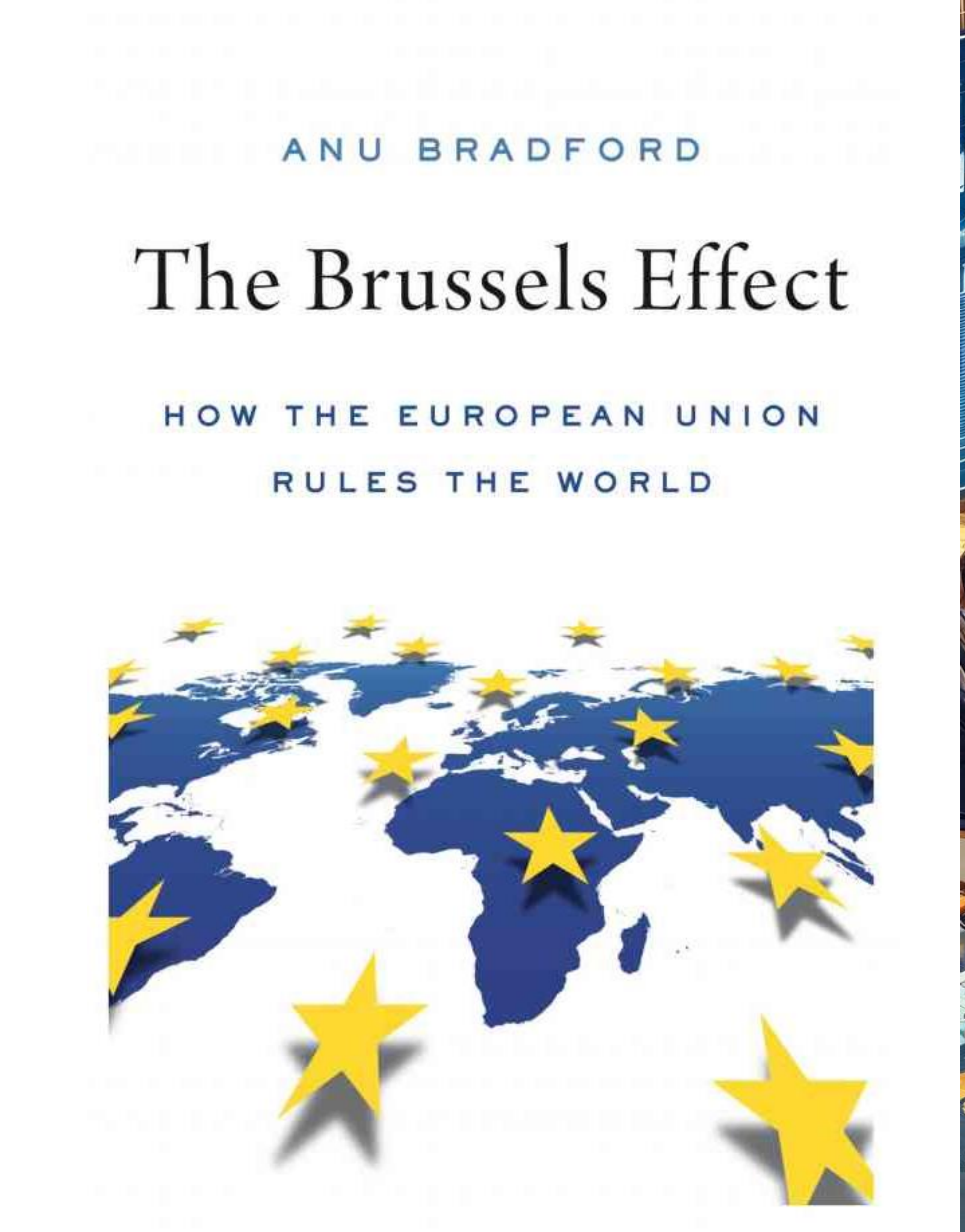
- Style transfer
- Machine translation
- Structured data to text data
- Rewriting data



GPT-NL

Training pipeline





ANU BRADFORD

The Brussels Effect

HOW THE EUROPEAN UNION RULES THE WORLD

What’s next?

- GPT-NL is made for and by the Netherlands
- To make it as relevant and useful as possible, your data and input is crucial

Want to participate? We need

- **Use Case providers**
- **Data providers**
- **End users**

More info or contact? Go to www.gpt-nl.nl or mail info@gpt-nl.nl