GPT-NL

FACILITEIT VOOR EEN SOEVEREIN NEDERLANDS TAALMODEL

Symposium E-Discovery 2024

GPT-NL team

TNO innovation for life

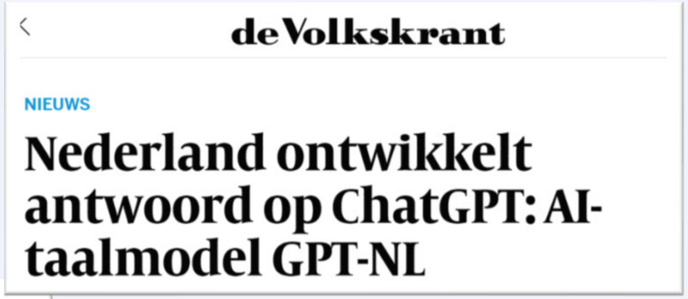
Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

SURF

GPT-NL Motivation

Why a Dutch LLM from scratch?

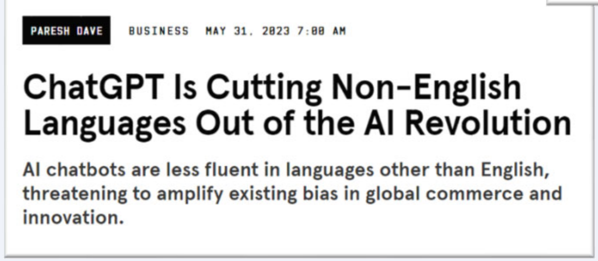
- Many of the current language models are trained on datasets that contain **no or very little Dutch data**
- **European values around bias, inclusivity and explainability** are insufficiently guaranteed in current solutions
- **Digital sovereignty** of European language and speech technology, no dependence on foreign multinationals
- **Privacy and IP**



de Volkskrant

NIEUWS

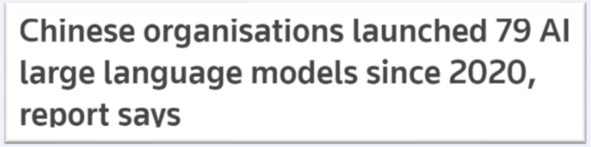
Nederland ontwikkelt antwoord op ChatGPT: AI-taalmodel GPT-NL



PARESH DAVE BUSINESS MAY 31, 2023 7:00 AM

ChatGPT Is Cutting Non-English Languages Out of the AI Revolution

AI chatbots are less fluent in languages other than English, threatening to amplify existing bias in global commerce and innovation.



Chinese organisations launched 79 AI large language models since 2020, report says

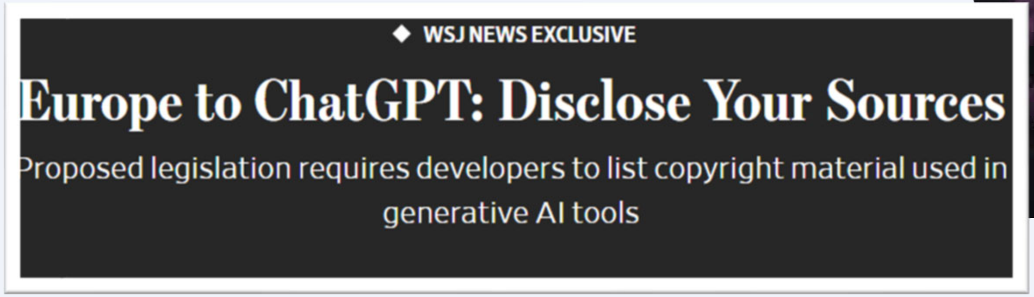
The Economist
India: irascible and indispensable
America's new industrial geography
Is the housing slump over?
Undoing business in China
BritGPT
How to make Britain an AI superpower

Große KI-Modelle

FÜR DEUTSCHLAND

Machbarkeitsstudie 2023

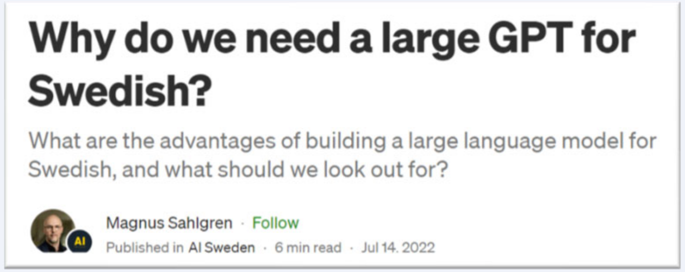
LEAM:AI KI BUNDESVERBAND



◆ WSJ NEWS EXCLUSIVE

Europe to ChatGPT: Disclose Your Sources

Proposed legislation requires developers to list copyright material used in generative AI tools



Why do we need a large GPT for Swedish?

What are the advantages of building a large language model for Swedish, and what should we look out for?

Magnus Sahlgren · Follow

Published in AI Sweden · 6 min read · Jul 14, 2022

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

WHAT?

we will build our own Dutch-English (50%-50%) language models from scratch,

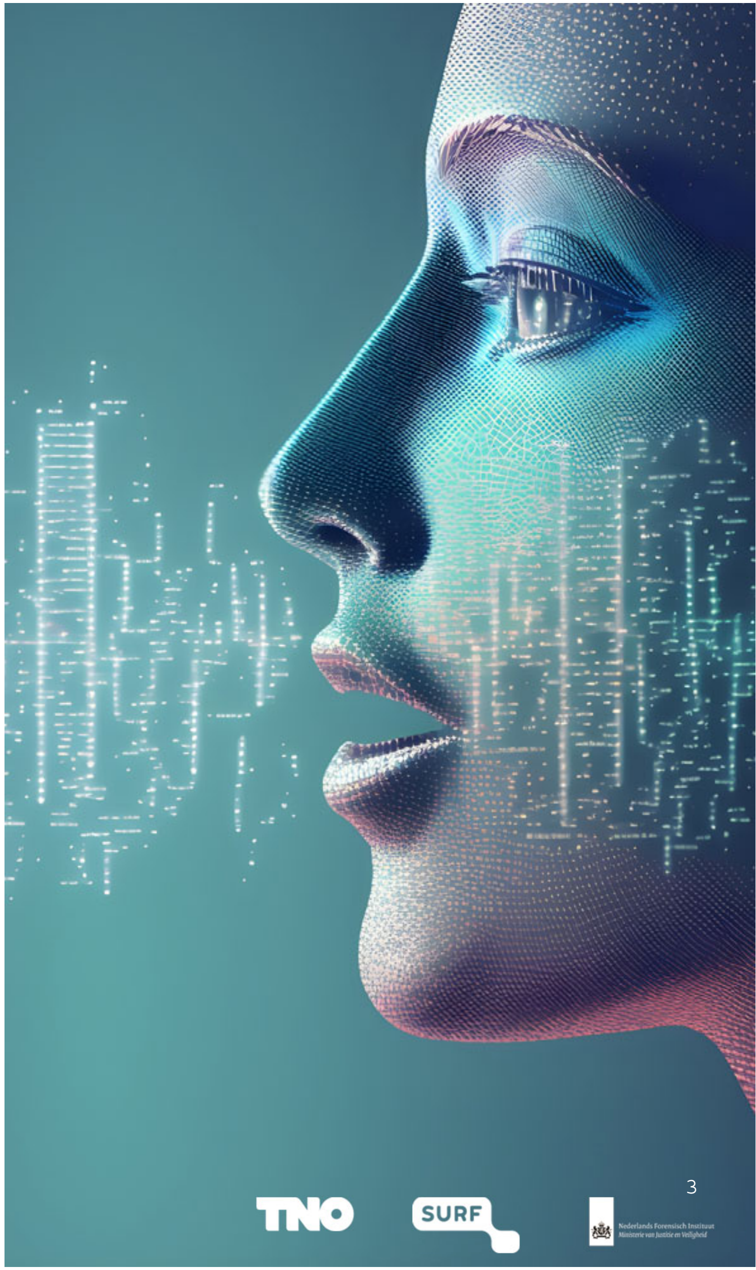
using data that we are allowed to use,
with privacy information removed,
with full transparency in our choices

Where we strife to be as transparent and compliant as possible

Small and large trained language models

On-premise fine-tuning cluster

Open code


TNO
SURF
Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

As transparent as possible

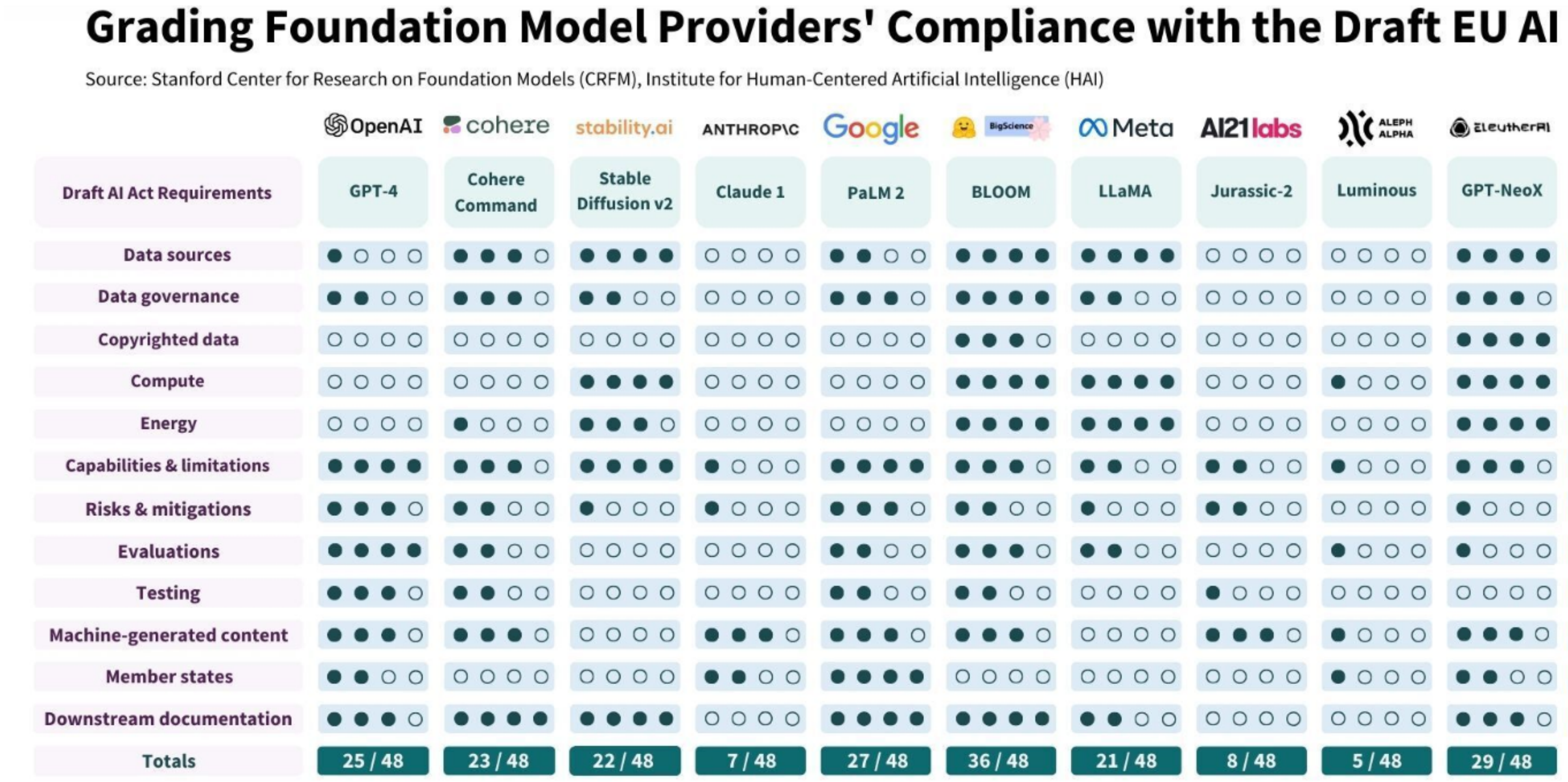
Minimal set of commitments for Responsible AI development:

- Have clear rules of engagement and communicate at regular intervals.
- Publish a **decision workflow document** to support dataset building.
- Publish a **definition of success** (both technical and societal benchmarks).
- Announce **stakeholder consultation opportunities** with fixed time windows.
- Report on ethical dilemmas and decisions as part of the base **reporting** process.
- **Open source code**: All code will be published.
- Publish **dataset- and model-cards** according to industry best practices.
- Review commitments on a regular basis to incorporate broad feedback.

(Public) commitment to responsibility ambitions, helps us keep ourselves accountable.

Ensuring auditability

As compliant as possible

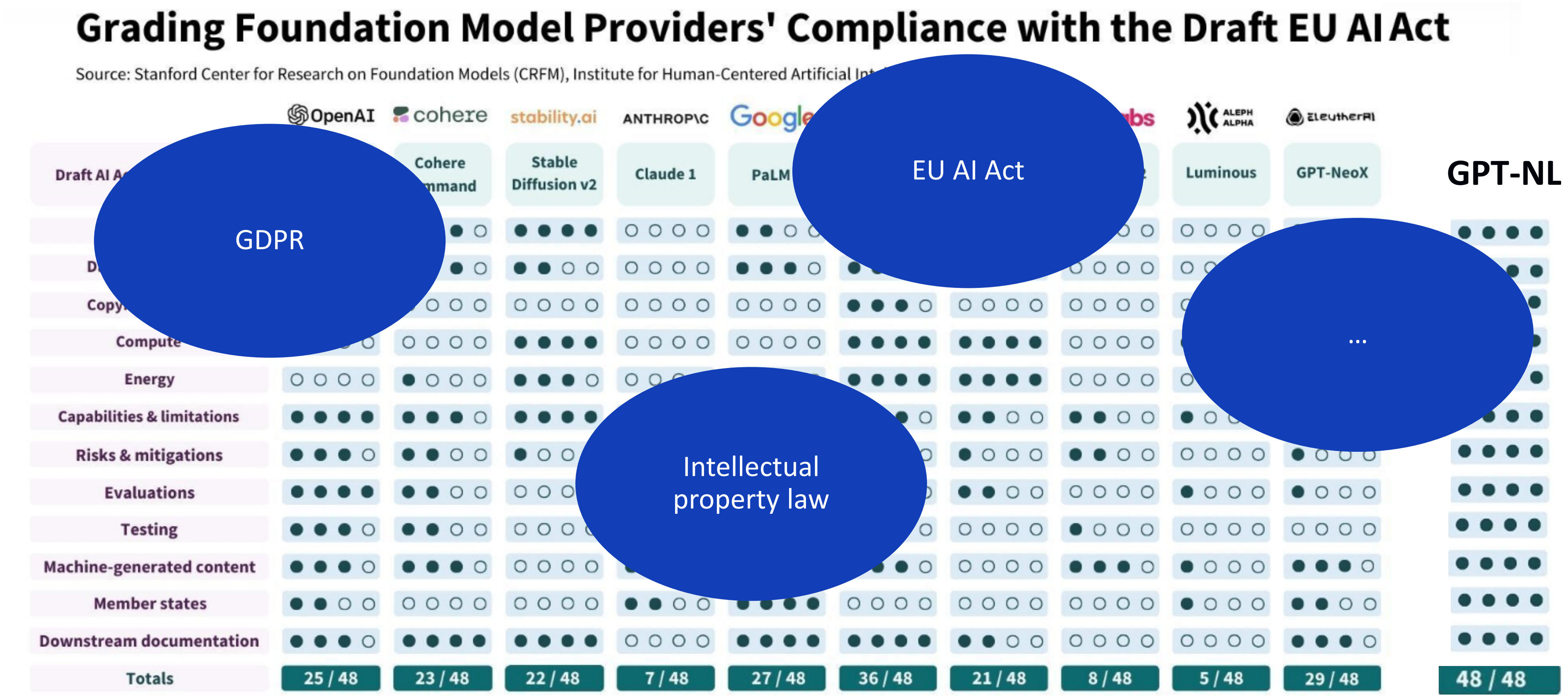


Grading Foundation Model Providers' Compliance with the Draft EU AI

Source: Stanford Center for Research on Foundation Models (CRFM), Institute for Human-Centered Artificial Intelligence (HAI)

	OpenAI	cohere	stability.ai	ANTHROPIC	Google	BigScience	Meta	AI21labs	ALEPH ALPHA	EleutherAI
Draft AI Act Requirements	GPT-4	Cohere Command	Stable Diffusion v2	Claude 1	PaLM 2	BLOOM	LLaMA	Jurassic-2	Luminous	GPT-NeoX
Data sources	●○○○	●●●○	●●●●	○○○○	●●○○	●●●●	●●●●	○○○○	○○○○	●●●●
Data governance	●●○○	●●●○	●●○○	○○○○	●●●○	●●●●	●●○○	○○○○	○○○○	●●●○
Copyrighted data	○○○○	○○○○	○○○○	○○○○	○○○○	●●●○	○○○○	○○○○	○○○○	●●●●
Compute	○○○○	○○○○	●●●●	○○○○	○○○○	●●●●	●●●●	○○○○	●○○○	●●●●
Energy	○○○○	●○○○	●●●○	○○○○	○○○○	●●●●	●●●●	○○○○	○○○○	●●●●
Capabilities & limitations	●●●●	●●●○	●●●●	●○○○	●●●●	●●●○	●●○○	●●○○	●○○○	●●●○
Risks & mitigations	●●●○	●●○○	●○○○	●○○○	●●●○	●●○○	●○○○	●●○○	○○○○	●○○○
Evaluations	●●●●	●●○○	○○○○	○○○○	●●○○	●●●○	●●○○	○○○○	●○○○	●○○○
Testing	●●●○	●●○○	○○○○	○○○○	●●○○	●●○○	○○○○	●○○○	○○○○	○○○○
Machine-generated content	●●●○	●●●○	○○○○	●●●○	●●●○	●●●○	○○○○	●●●○	●○○○	●●●○
Member states	●●○○	○○○○	○○○○	●●○○	●●●●	○○○○	○○○○	○○○○	●○○○	●●○○
Downstream documentation	●●●○	●●●●	●●●●	○○○○	●●●●	●●●●	●●○○	○○○○	○○○○	●●●○
Totals	25 / 48	23 / 48	22 / 48	7 / 48	27 / 48	36 / 48	21 / 48	8 / 48	5 / 48	29 / 48

As compliant as possible



Grading Foundation Model Providers' Compliance with the Draft EU AI Act

Source: Stanford Center for Research on Foundation Models (CRFM), Institute for Human-Centered Artificial Int

	OpenAI	cohere	stability.ai	ANTHROP\C	Google			bs	ALEPH ALPHA	EleutherAI	
Draft AI A		Cohere mmand	Stable Diffusion v2	Claude 1	PaLM			2	Luminous	GPT-NeoX	GPT-NL
		● ○	● ● ● ●	○ ○ ○ ○	● ● ○ ○			○ ○	○ ○ ○ ○		● ● ● ●
D		● ○	● ● ○ ○	○ ○ ○ ○	● ● ● ○	●		○ ○ ○ ○	○ ○		● ●
Copy		○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ○	○ ○ ○ ○	○ ○ ○ ○	○		●
Compute	○	○ ○ ○ ○	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	●		●
Energy	○ ○ ○ ○	● ○ ○ ○	● ● ● ○	○ ○		● ● ● ●	● ● ● ●	○ ○ ○ ○	○		●
Capabilities & limitations	● ● ● ●	● ● ● ○	● ● ● ●			● ○	● ● ○ ○	● ● ○ ○	● ○		● ● ●
Risks & mitigations	● ● ● ○	● ● ○ ○	● ○ ○			○	● ○ ○ ○	● ● ○ ○	○ ○ ○ ○	● ○ ○ ○	● ● ● ●
Evaluations	● ● ● ●	● ● ○ ○	○ ○ ○			○	● ● ○ ○	○ ○ ○ ○	● ○ ○ ○	● ○ ○ ○	● ● ● ●
Testing	● ● ● ○	● ● ○ ○	○ ○ ○			○	○ ○ ○ ○	● ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●
Machine-generated content	● ● ● ○	● ● ● ○	○ ○ ○ ○	●		● ● ○	○ ○ ○ ○	● ● ● ○	● ○ ○ ○	● ● ● ○	● ● ● ●
Member states	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ○ ○	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ○ ○ ○	● ● ○ ○	● ● ● ●
Downstream documentation	● ● ● ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	● ● ● ●	● ● ● ●	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ○	● ● ● ●
Totals	25 / 48	23 / 48	22 / 48	7 / 48	27 / 48	36 / 48	21 / 48	8 / 48	5 / 48	29 / 48	48 / 48

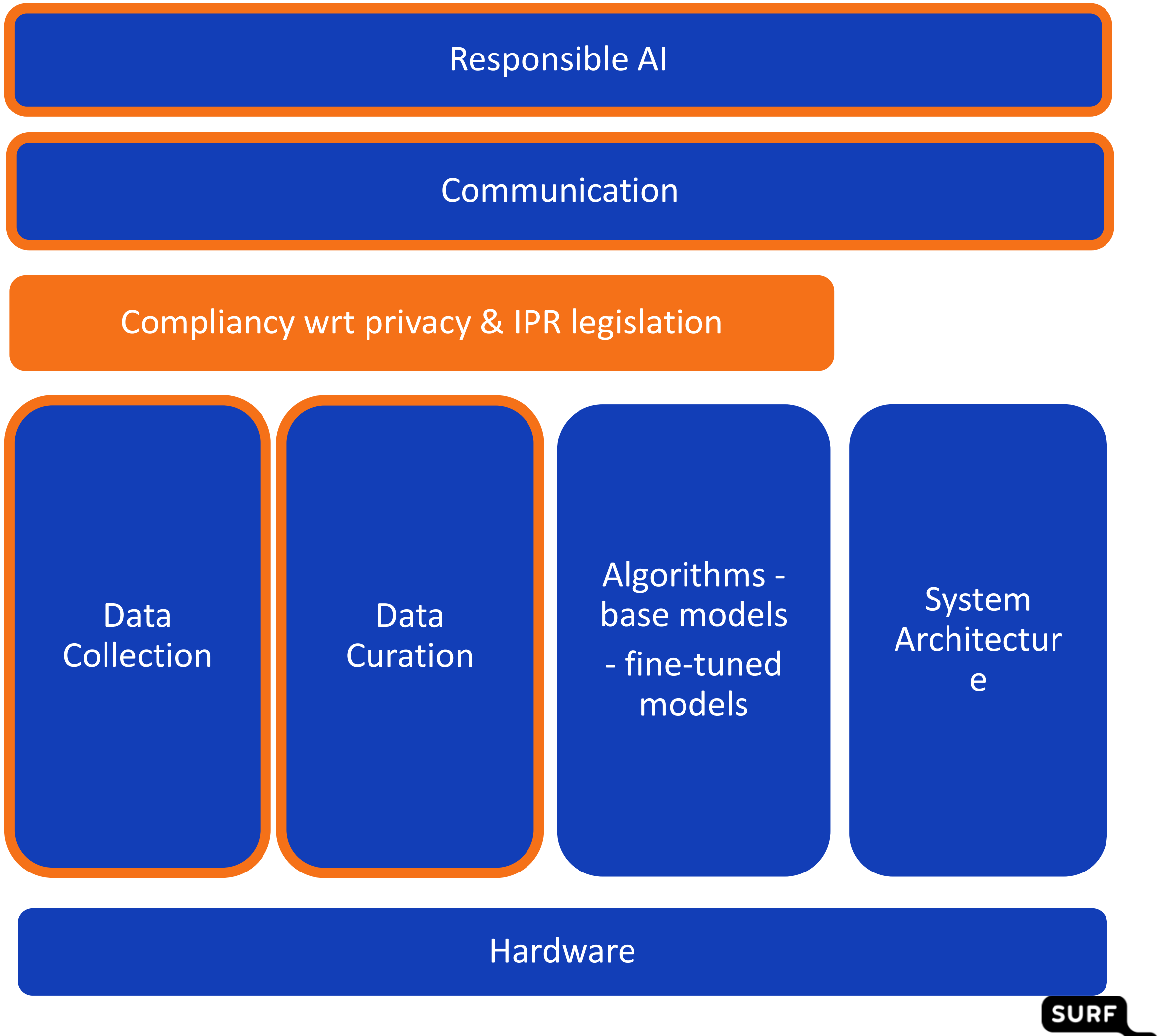
GDPR

EU AI Act

Intellectual property law

...

HOW ?



Responsible AI

Communication

Compliancy wrt privacy & IPR legislation

Data Collection

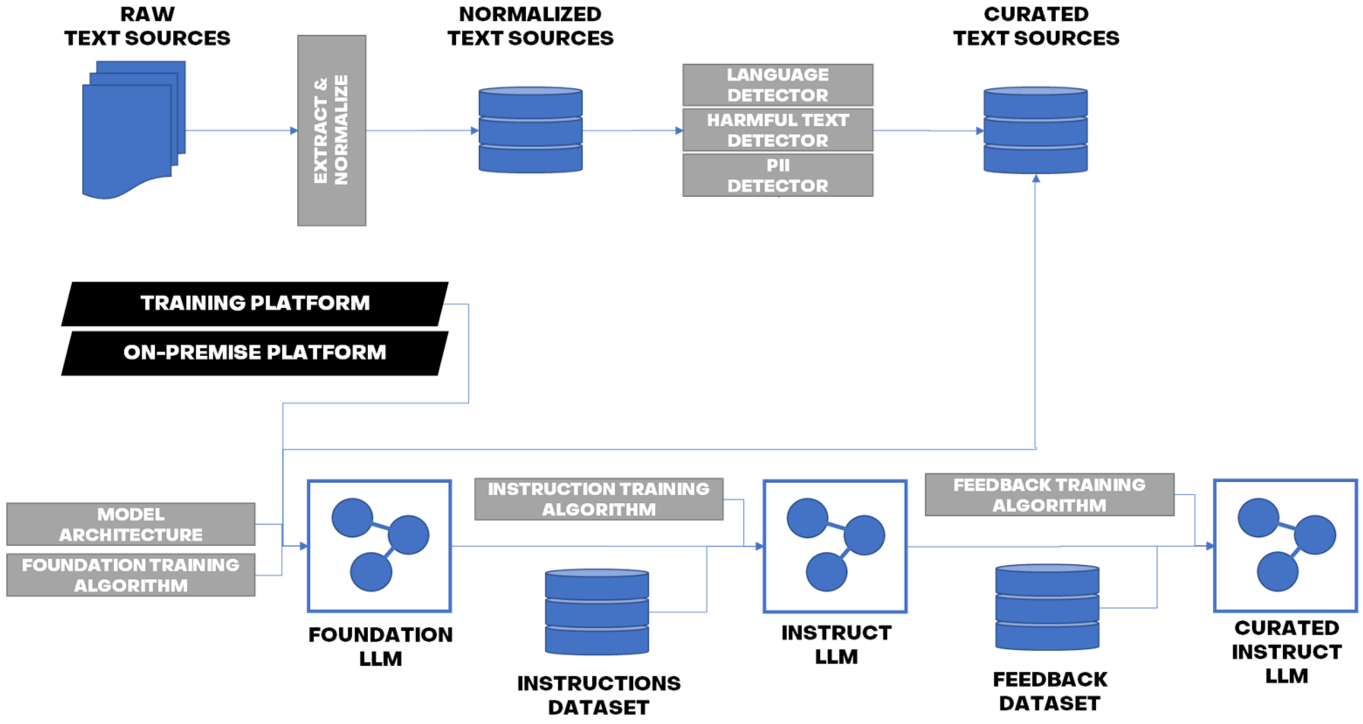
Data Curation

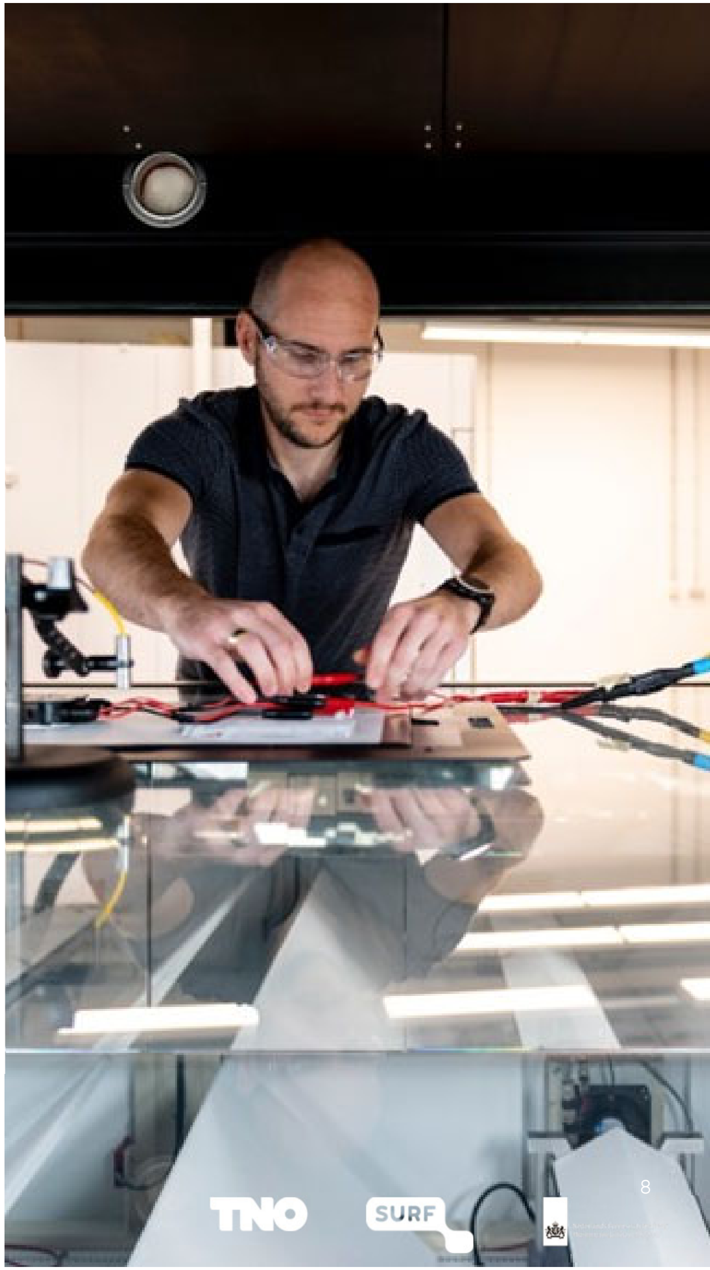
Algorithms - base models - fine-tuned models

System Architectur e

Hardware

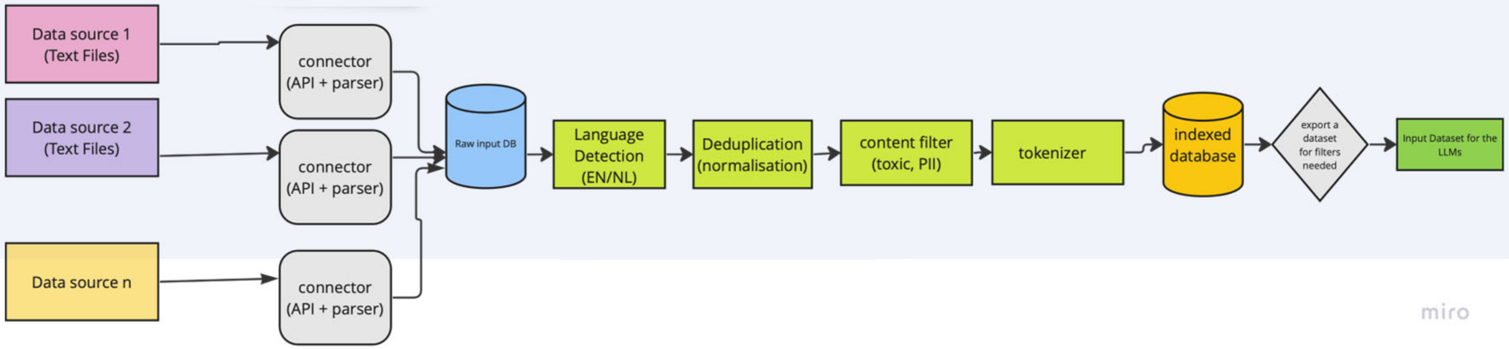
HOW ?


RAW TEXT SOURCES
EXTRACT & NORMALIZE
NORMALIZED TEXT SOURCES
LANGUAGE DETECTOR
HARMFUL TEXT DETECTOR
PII DETECTOR
CURATED TEXT SOURCES
TRAINING PLATFORM
ON-PREMISE PLATFORM
MODEL ARCHITECTURE
FOUNDATION TRAINING ALGORITHM
FOUNDATION LLM
INSTRUCTION TRAINING ALGORITHM
INSTRUCTIONS DATASET
INSTRUCT LLM
FEEDBACK TRAINING ALGORITHM
FEEDBACK DATASET
CURATED INSTRUCT LLM

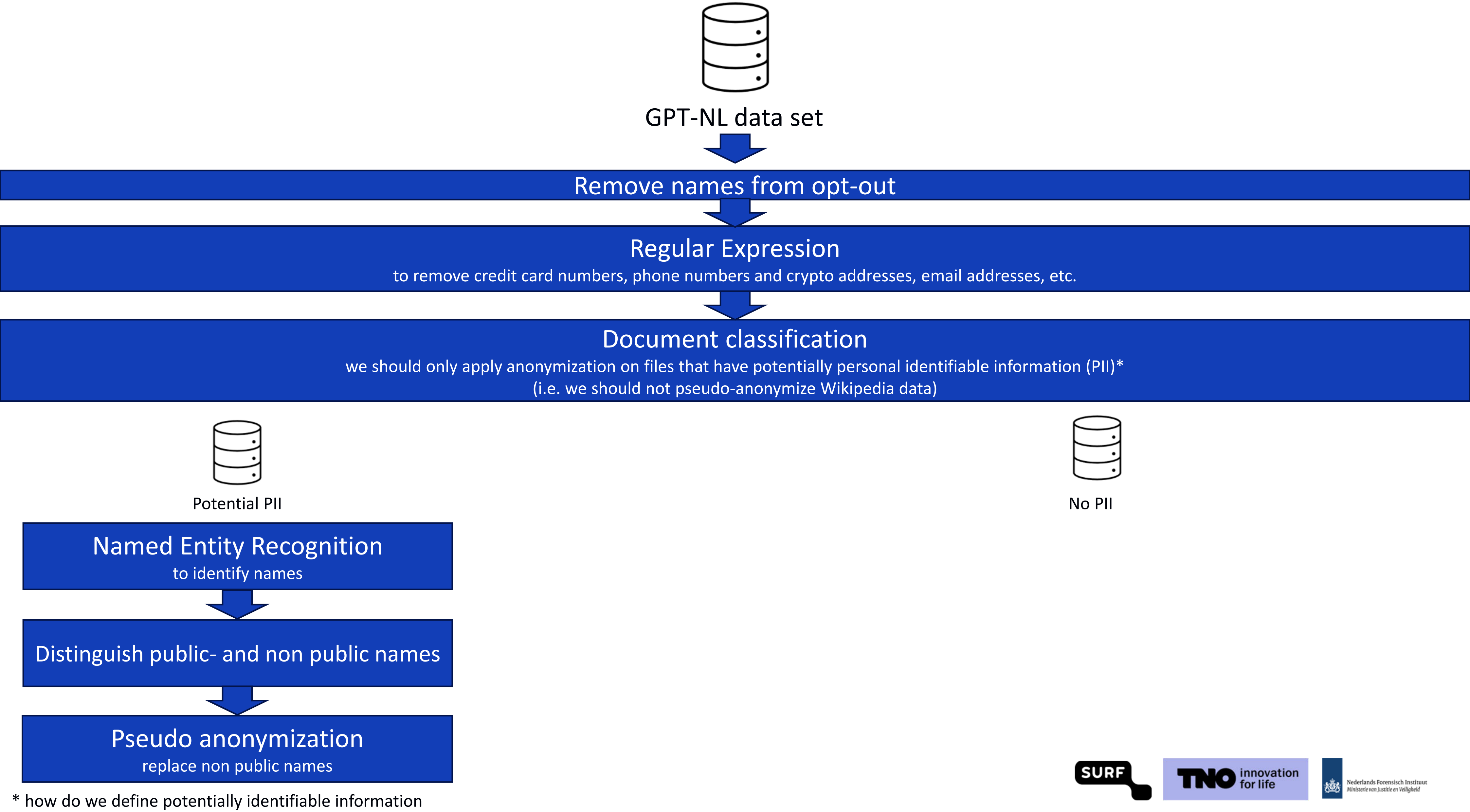


Data

- Texts from sources with permissive licenses (CC-0, …) and based on agreements with data holders
- Aim: at least 500B Dutch tokens
- Data curation absolutely crucial
- Language detection, harmful text detection, bias detection, deduplication and PII detection and removal


Data source 1 (Text Files)
Data source 2 (Text Files)
Data source n
connector (API + parser)
connector (API + parser)
connector (API + parser)
Raw input DB
Language Detection (EN/NL)
Deduplication (normalisation)
content filter (toxic, PII)
tokenizer
indexed database
export a dataset for filters needed
Input Dataset for the LLMs
miro

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



GPT-NL data set

Remove names from opt-out

Regular Expression

to remove credit card numbers, phone numbers and crypto addresses, email addresses, etc.

Document classification

we should only apply anonymization on files that have potentially personal identifiable information (PII)*
(i.e. we should not pseudo-anonymize Wikipedia data)

Potential PII

No PII

Named Entity Recognition

to identify names

Distinguish public- and non public names

Pseudo anonymization

replace non public names

* how do we define potentially identifiable information

SURF

TNO innovation for life

Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

WHEN?

Q4 2023 – september 2024

Build up & curate data

- License agreement dataholders, DPIAs & DPAs
- Select Dutch & English materials
- modules to filter and process data

October – Q2 2025

Training language models

- Foundation models
- Fine-tuning rounds

European embedding

- Good connections with similar initiatives in other European countries
  - OpenGPT-X
  - AI Sweden
  - Silo.ai
  - Catalan initiatives
  - Alliance for Language Technolgies (ALT) EDIC
  - Language Data Spaces
  - ..

What’s next?

- GPT-NL is voor en door Nederland gemaakt
- Om het zo relevant mogelijk te maken, is jullie stem cruciaal

Meedoen? Je kan meedoen als

- **Use Case provider**
- **Data provider**
- **End user**

Contact? Mail saskia.lensink@tno.nl