Ontwikkelingen en veiligheid rondom

Saskia Lensink & Joachim de Greeff

NLAIC congres, 24 April 2024

TNO innovation for life

Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

SURF

Who are we?

- Saskia works as a Business Consultant with the TNO Data Science group and is active with the Dutch Language Speech Coalition
- She is a linguist and applies her knowledge in the field of AI, Large Language Models, and conversational AI across diverse projects
- Additionally, she is active in a broad range of consortia and networks to promote sovereign and high-quality language models for and by Europe
- Currently she is the TNO product owner for GPT-NL

- Joachim works as deputy research manager of the TNO Data Science group
- Previously he was lead of the Communicative AI topic group, which aims for seamless and natural information exchange between humans and AI systems
- He has a background in human-robot interaction and has spent a number of years in the academic world working on the interplay between humans and AI
- Joachim's passion is to make AI systems more socially aware, so that interaction with humans is as intuitive and natural as possible

Starting note Nederlandse AI voor het Nederlands (NAIN)

Versie: 0.9, 5 mei 2020

Auteurs
Erwin van der Eijk, NFI
Lisanne van Dijk, NFI
Frans Nauta, Data Science Initiative
Sander Ruiter, Nederlandse AI Coalitie

Een grote belemmering voor de benutting van AI in Nederland is dat bestaande algoritmen niet goed getraind zijn op de Nederlandse taal. [...] Dit probleem [...] geldt voor de gehele publieke sector en voor alle Nederlandstalige interacties in de markt.

Individuele organisaties ontwikkelen soms deeloplossingen voor specifieke domeinen, maar zonder een overkoepelend idee omdat daarvoor onvoldoende geld is. Om die reden ziet de werkgroep veiligheid het NAIN als een flagship voor de NLAIC.

Starting note Nederlandse AI voor het Nederlands (NAIN)

Versie: 0.9, 5 mei 2020

Auteurs
Erwin van der Eijk, NFI
Lisanne van Dijk, NFI
Frans Nauta, Data Science Initiative
Sander Ruiter, Nederlandse AI Coalitie

[...] de resultaten van dit project [zijn] ***overal in de Nederlandse samenleving bruikbaar****, iedere dag, honderden miljoenen keren. Bij het registreren van zorghandeling zonder dat een zorgverlener haar handen van het bed hoeft te halen, bij het bellen van 112 om een ongeluk te melden, op de Twitter-feed van KLM om klanten snel en adequaat te woord te staan, voor het analyseren van terroristische dreigingen op een WhatsApp chat, automatische (en betrouwbare) ondertiteling van Nederlands beeld- en audiomateriaal, automatische transcripties van pathologisch onderzoek, van vergaderingen, enz.*

[Dit project] maakt een enorme diversiteit aan toepassingen mogelijk [...], met grote publieke en economische waarde.

HSD
securitydelta.nl

TNO innovation for life

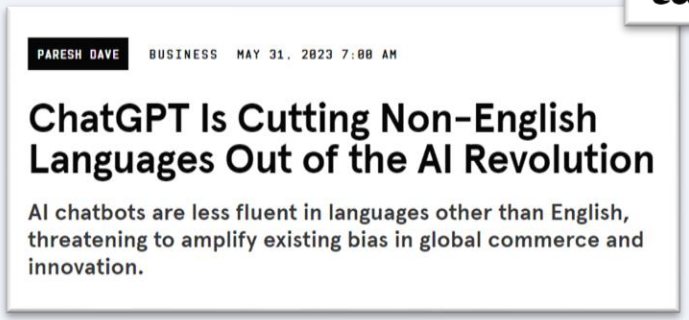
Why a Dutch LLM from scratch?

- 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


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

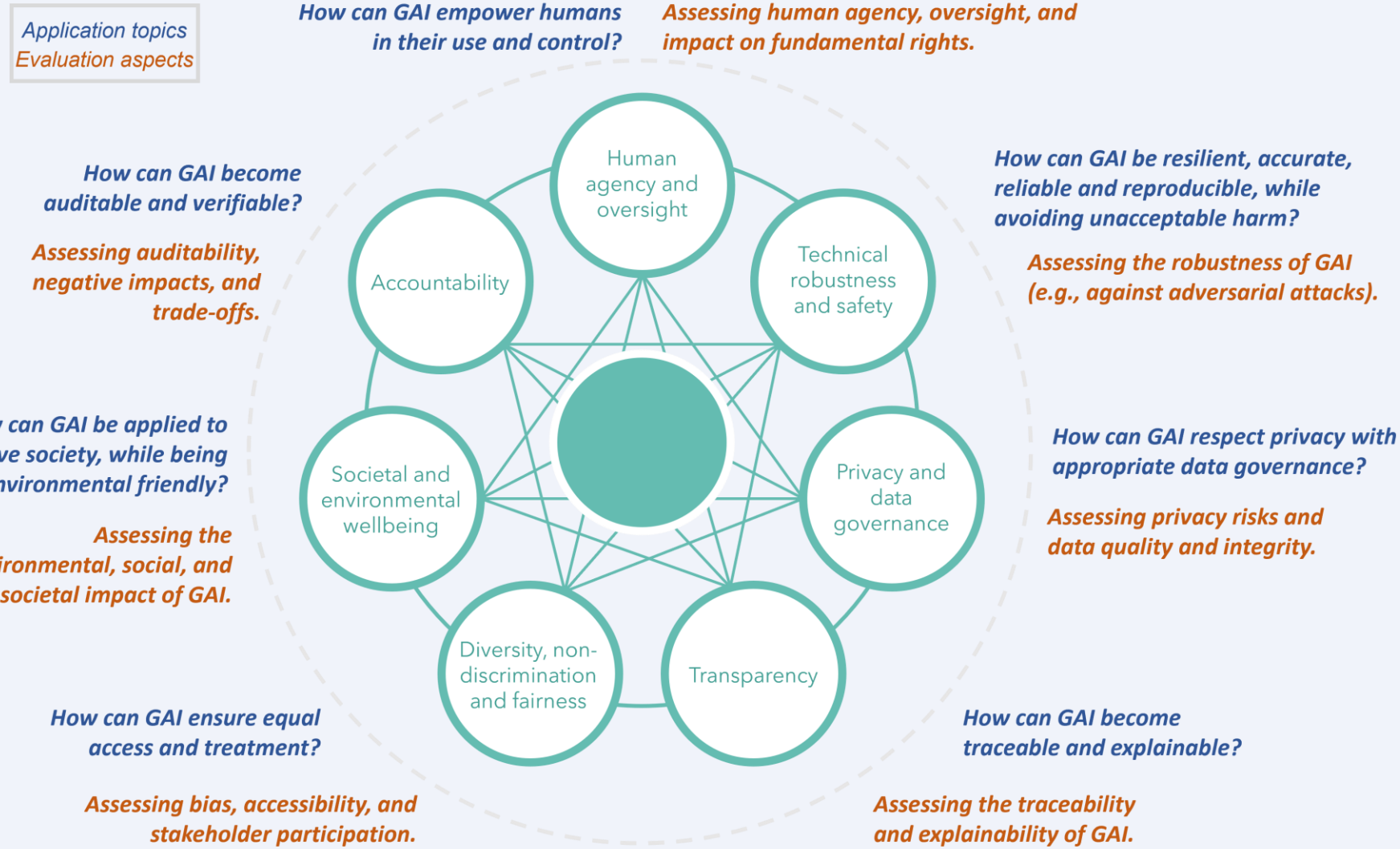
WHY?

FOUNDATION MODEL	INSTRUCT MODEL		
		limited availability for academic use	
limited availability for academic use		limited fine-tuning	limited on-premise availability
limited fine-tuning		English style texts	API
RAW TEXT DATA	**INSTRUCTIONS**	**FEEDBACK**	
mainly English	not public	not public	
contains copyrighted material		describe preferences of US big-tech	
contains names, email, phone			
kept secret			
no opt-out			

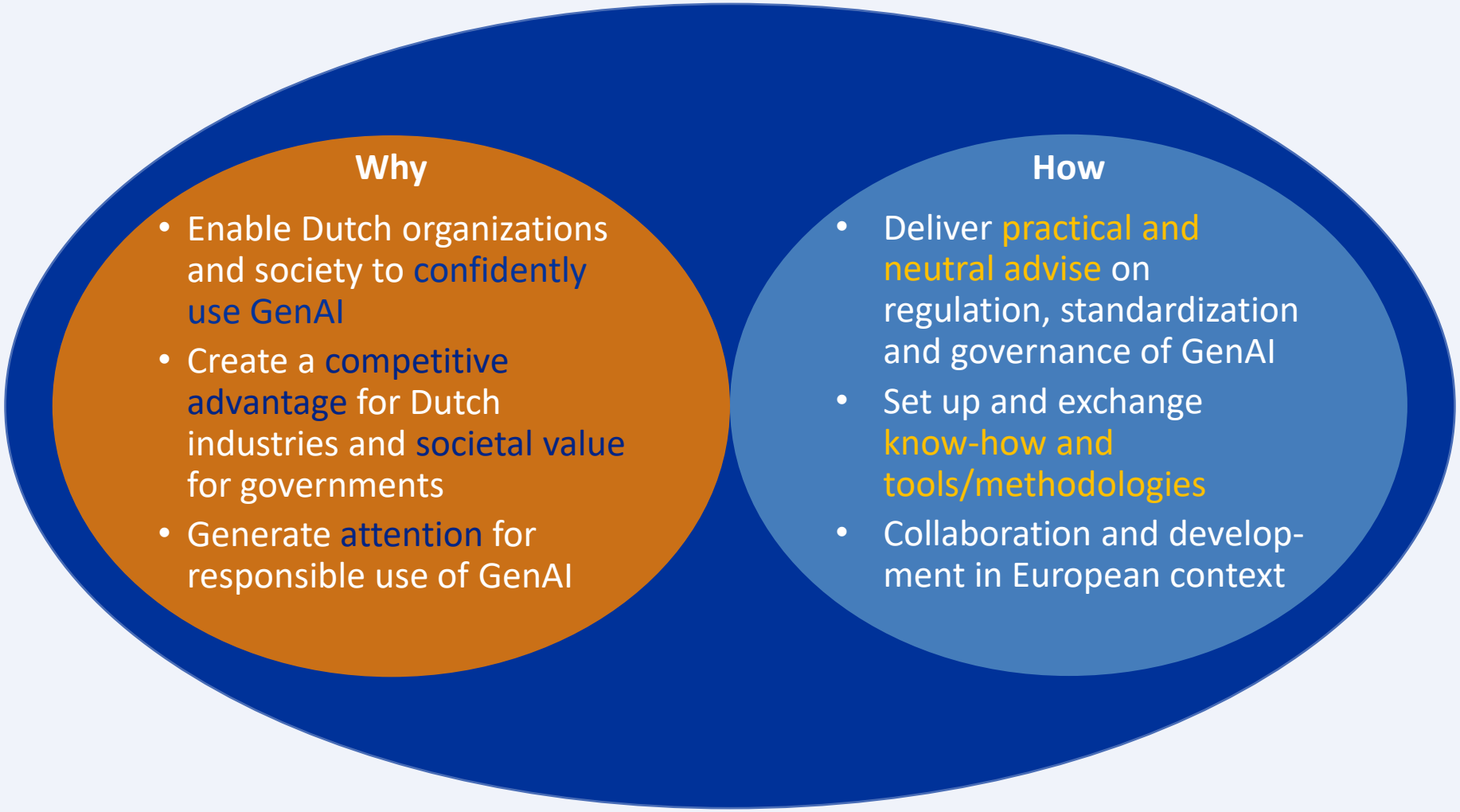
What do you think?

Stakeholders in the Ecosystem and their topics and evaluation aspects, derived from the EU requirements for trustworthy AI

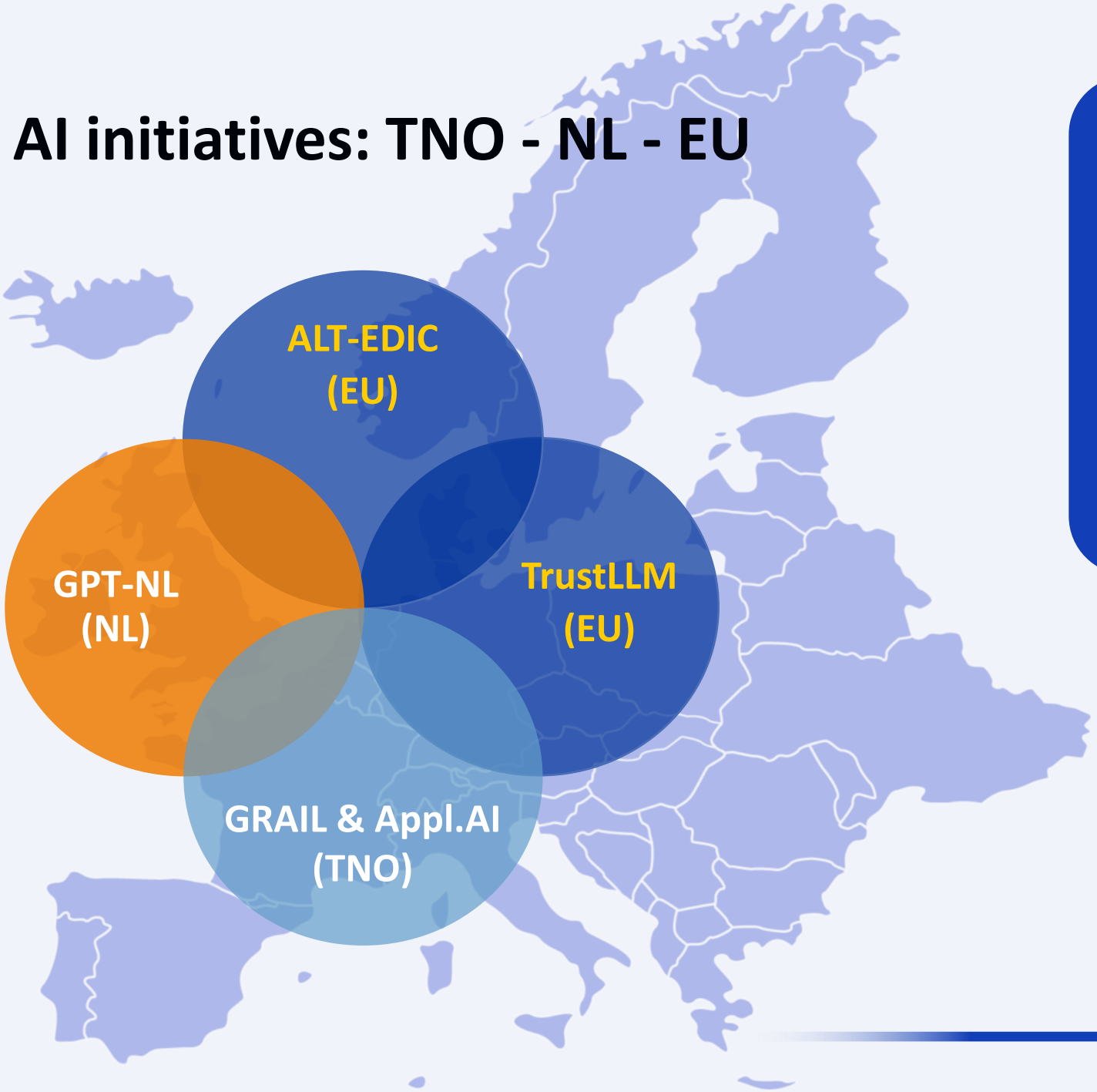
- Policy makers and Authorities
- GAI Users / Associations
- Research- and Knowledge Institutions
- GAI-/LLM-/ Platform-Providers and Integrators
- GAI Auditors & Consultancies



Objectives for an Ecosystem for Responsible Generative AI



(Generative) AI initiatives: TNO - NL - EU



EU Ecosystem Trustworthy Generative AI

OpenGPT-X

AI Sweden

Silo.ai

Catalan initiatives

Language Data Spaces

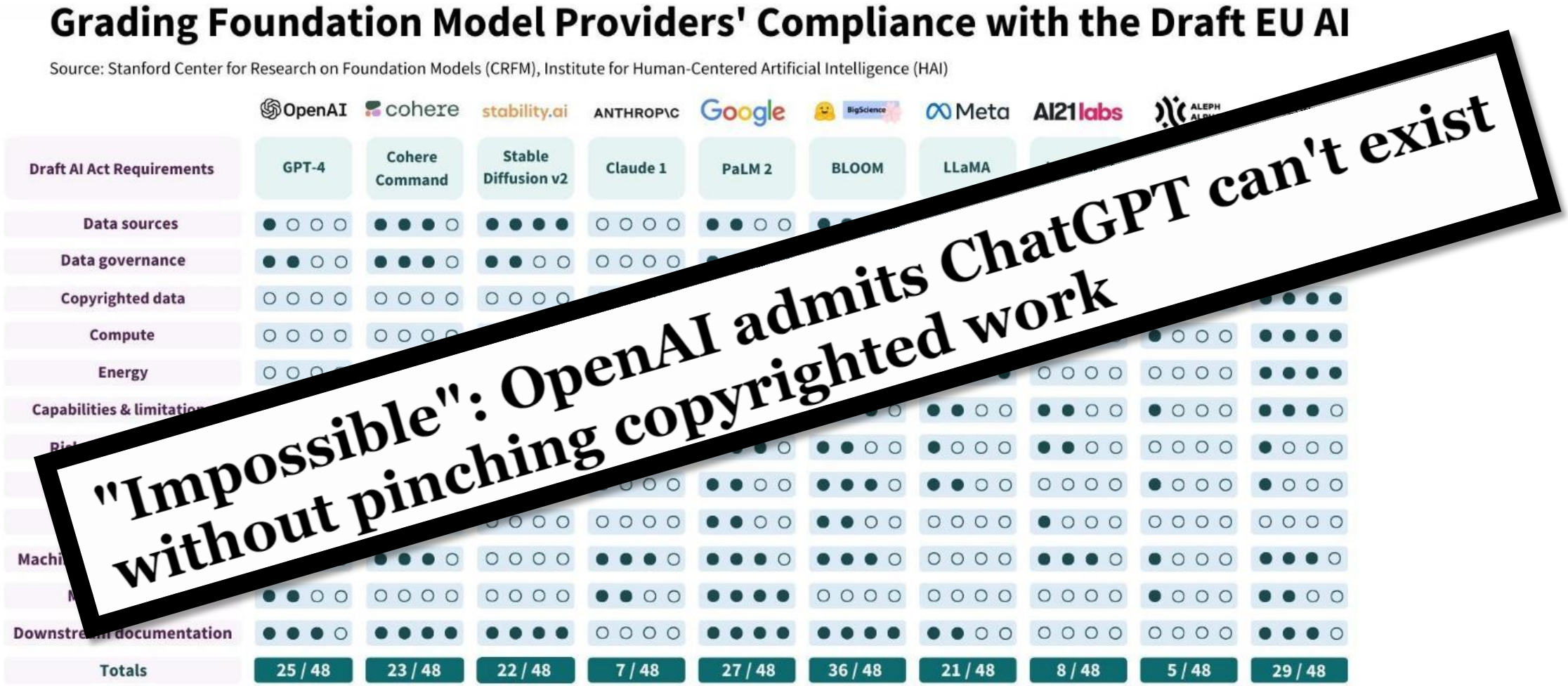
…

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)

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


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
Draft AI Act Requirements
GPT-4
Cohere Command
Stable Diffusion v2
Claude 1
PaLM 2
BLOOM
LLaMA
Data sources
Data governance
Copyrighted data
Compute
Energy
Totals
25 / 48
23 / 48
22 / 48
7 / 48
27 / 48
36 / 48
21 / 48
8 / 48
5 / 48
29 / 48
"Impossible": OpenAI admits ChatGPT can't exist without pinching copyrighted work

What do you think?

The GPT-NL consortium

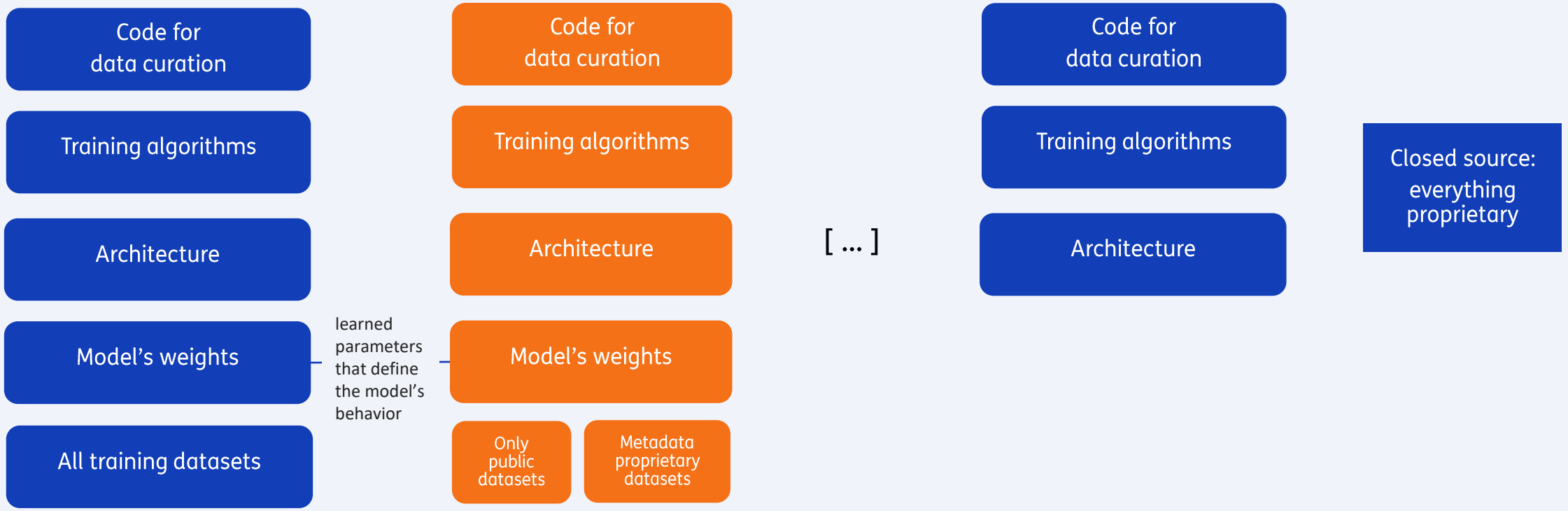
TNO innovation for life

SURF

Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

GPT-NL

Transparency code, model and data set



Complete transparency

Highest level of transparency, collaboration, and accessibility

Large limitations wrt size potential dataset

GPT-NL sweet spot

High level of transparency, collaboration and accessibility

Less limitations wrt size potential dataset

Less transparency

Less limitations wrt size potential dataset

More control, and less challenges related to quality control, resource allocation, and intellectual property management

No (responsible AI) audits possible

As transparent as possible

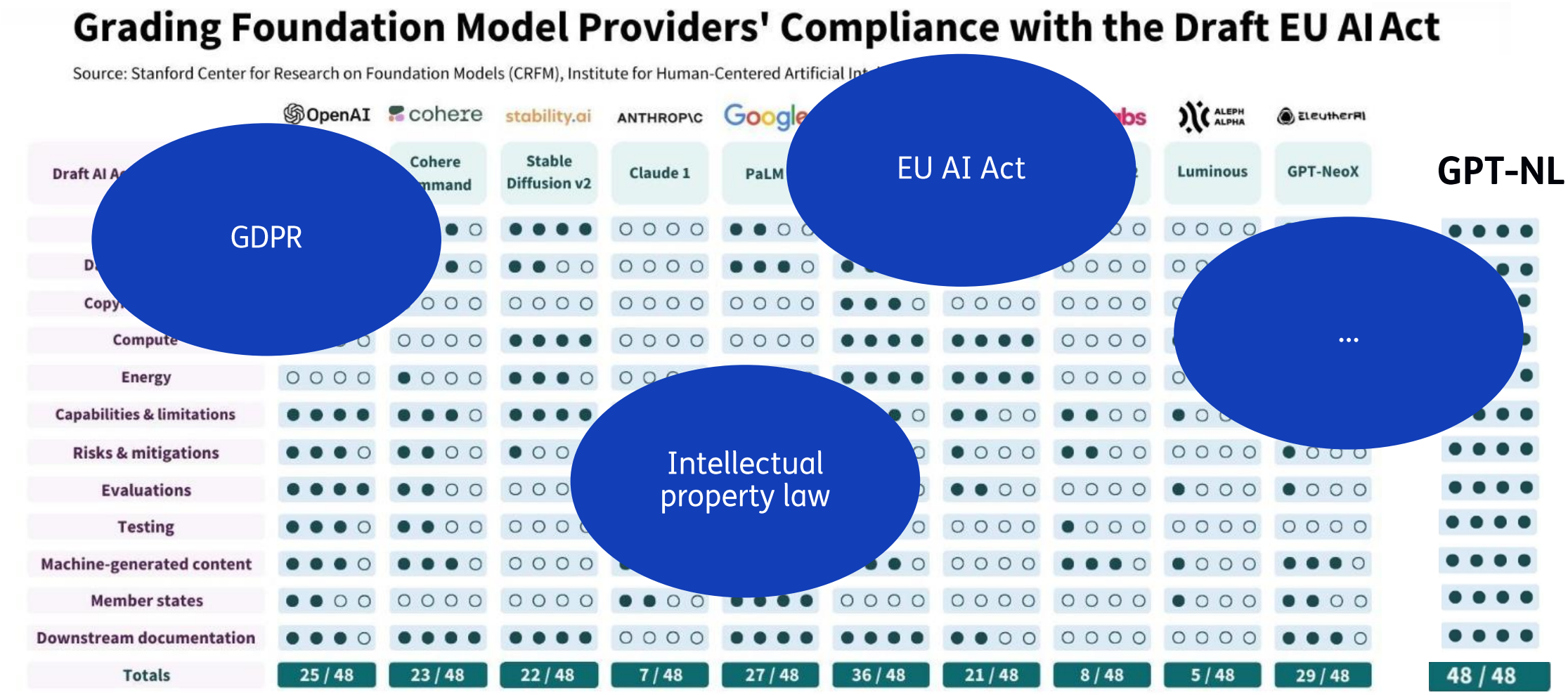
Minimal set of commitments for Responsible AI development:

- 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.

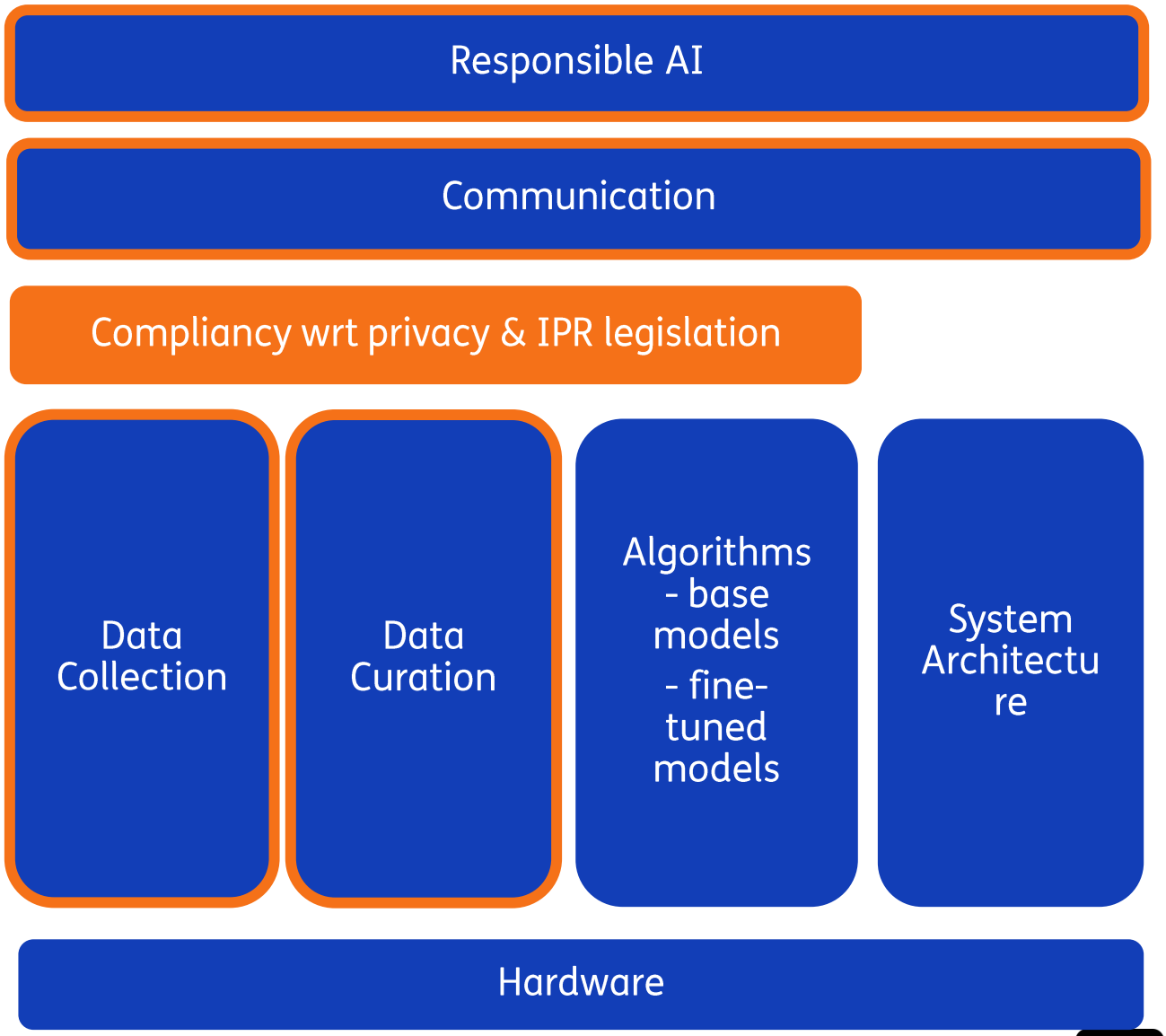
(Public) commitment to responsibility ambitions, helps us keep ourselves accountable.

Ensuring auditability

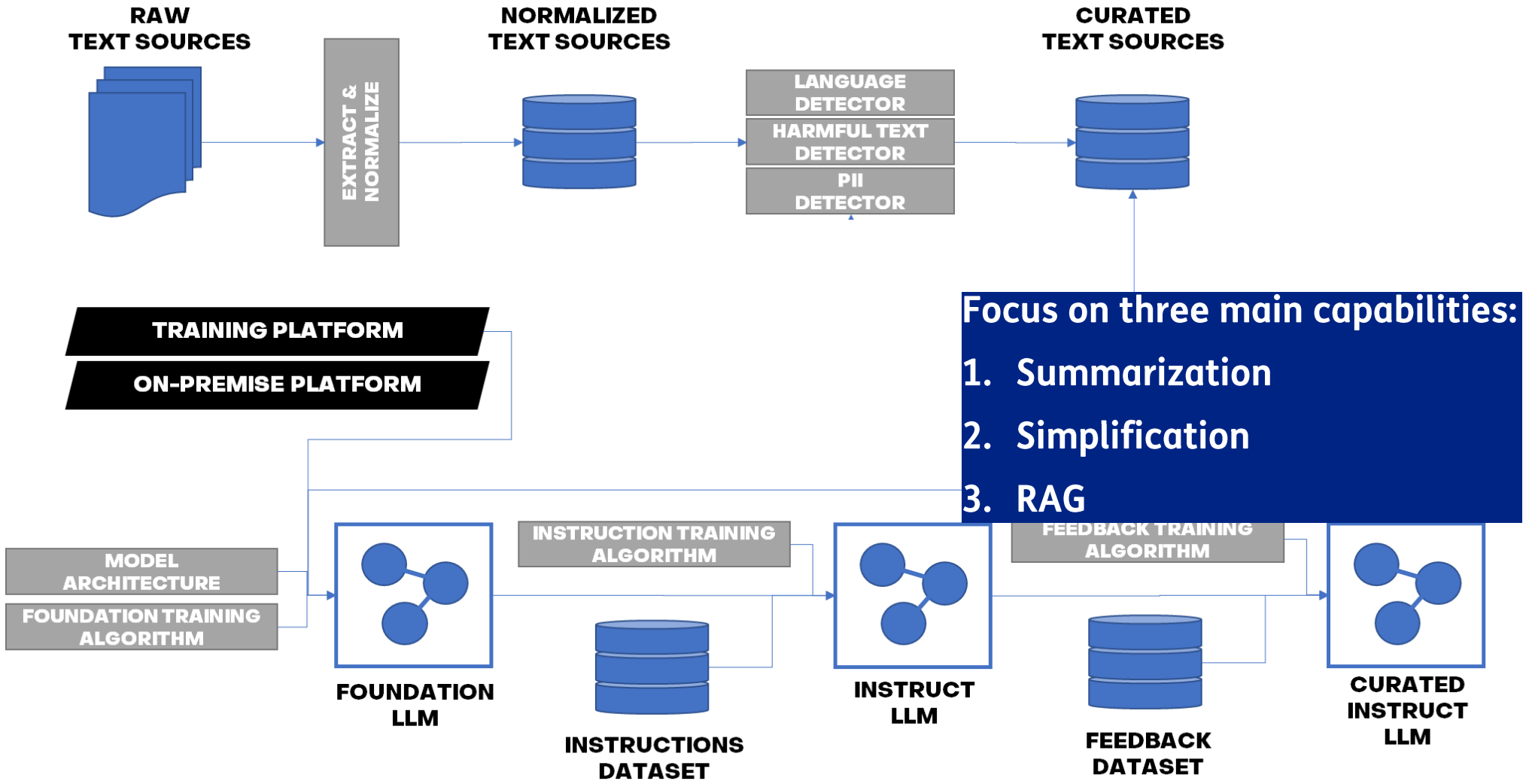
As compliant as possible



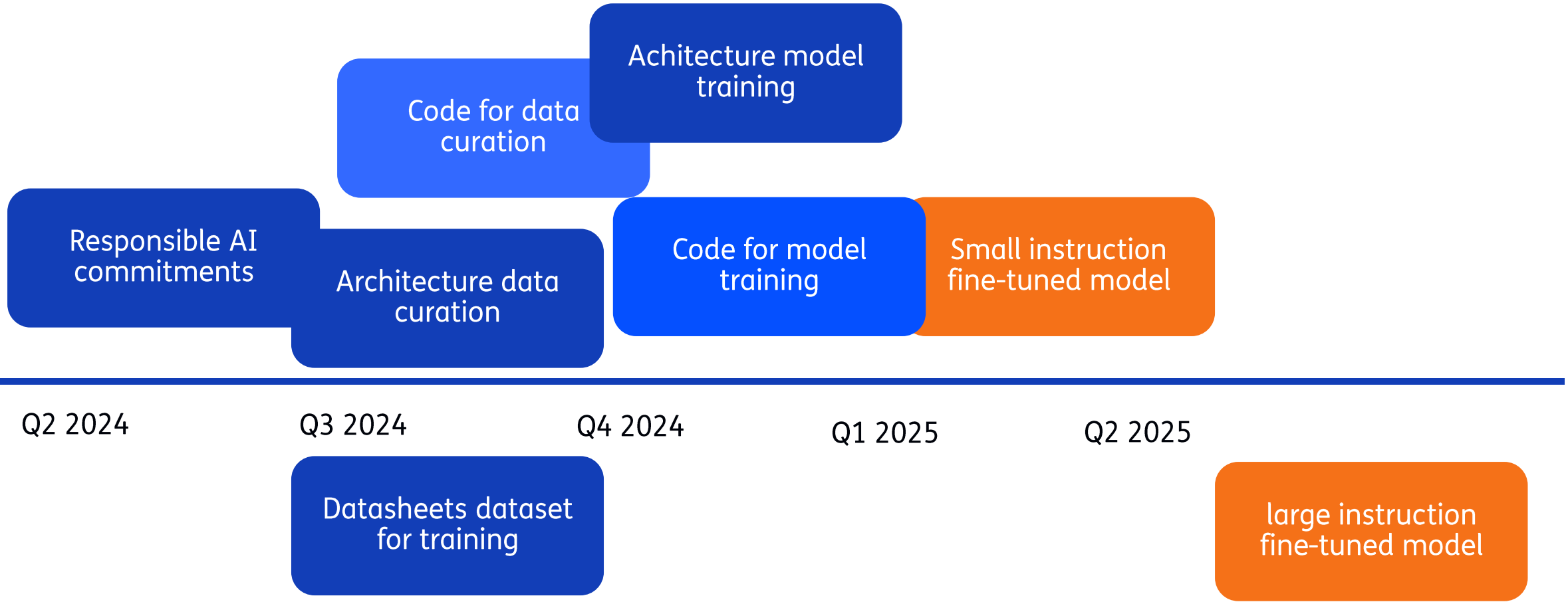
HOW?


Responsible AI
Communication
Compliancy wrt privacy & IPR legislation
Data Collection
Data Curation
Algorithms - base models - fine-tuned models
System Architectu re
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
Focus on three main capabilities:
1. Summarization
2. Simplification
3. RAG
MODEL ARCHITECTURE
FOUNDATION TRAINING ALGORITHM
FOUNDATION LLM
INSTRUCTION TRAINING ALGORITHM
INSTRUCTIONS DATASET
INSTRUCT LLM
FEEDBACK TRAINING ALGORITHM
FEEDBACK DATASET
CURATED INSTRUCT LLM

Roadmap


Achitecture model training
Code for data curation
Responsible AI commitments
Architecture data curation
Code for model training
Small instruction fine-tuned model
Q2 2024
Q3 2024
Q4 2024
Q1 2025
Q2 2025
Datasheets dataset for training
large instruction fine-tuned model
Project update
Project update
Project update
Project update
Project update

Ethical guidelines trustworthy AI

Feature	GPT-NL	NextGen genAI?	
Human agency & oversight	●●	●●●●●	
Technical robustness & safety	●●●	●●●●●	
Privacy and data governance	●●●●	●●●●●	
Transparency	●●●	●●●●●	
Diversity, non-discrimination and fairness	●●	●●●●●	
Accountability	●●●●	●●●●●	
Societal and environmental wellbeing	●●	●●●●●	

What do you think?

Potential use cases safety & security

	Summarization	Simplification	RAG
Compliance monitoring	New regulations, compliance requirements	Complex legal jargon	Access and integrate case law, precedents, protocols
Emergency response coordination	real-time updates from e.g. social media, emergency services, and news	complex emergency protocols and instructions into easy-to-follow steps for the general public	leveraging past data on similar emergency situations, suggest optimized response strategies
			dynamically update FAQs and provide real-time answers to public inquiries during safety crises, using the most current guidelines and data
Cybersecurity	security reports, threat intelligence feeds, and incident logs	technical language found in cybersecurity documentation and threat intelligence	enriched, context-aware answers to queries about (emerging) cyber threats

TNO innovation for life