

# Facetten van Responsible AI met GPT-NL

Lieke & Julio  
TNO / GPT-NL

(TNO, SURF & NFI)

**GPT-NL**





## **Lieke Dom**

- Consultant Responsible Innovation @ TNO
- Responsible AI & Communication at GPT-NL

## **Julio Oliveira Filho**

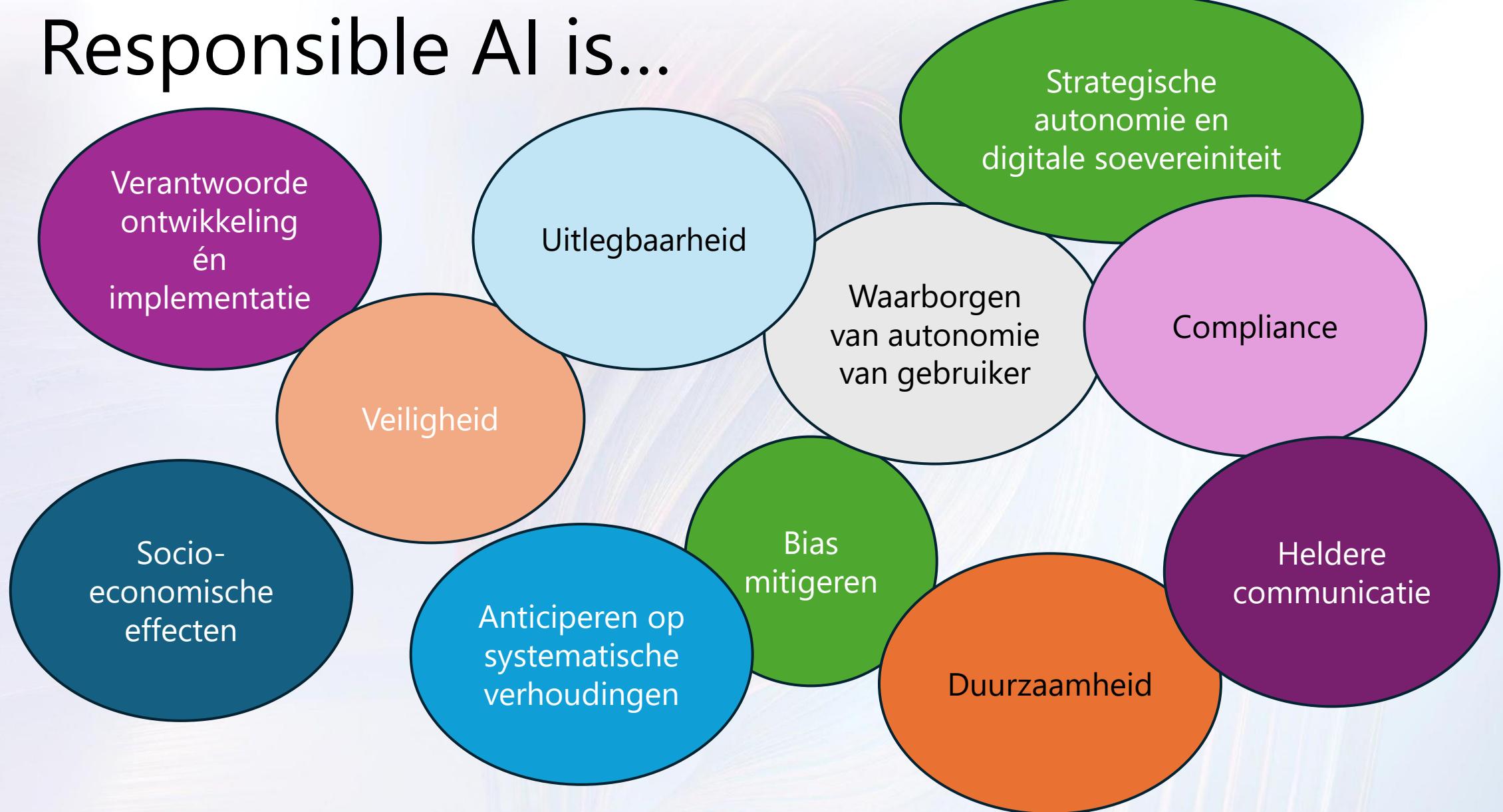
- Senior Innovation Scientist at the Advanced Computing Engineering Group at TNO
- Lead System Architect @ GPT-NL



# Waar denken jullie aan bij Responsible AI?

Deel 1 – Facetten van Responsible AI

# Responsible AI is...



# Responsible AI is...



# Responsible AI met GPT-NL

Deel 2 – Bijdrages en uitdagingen



# Het GPT-NL Consortium



**TNO** innovation  
for life

SURF



Nederlands Forensisch Instituut  
Ministerie van Justitie en Veiligheid



**GPT-NL**

NIEUWS

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

# Motivatie

Veel huidige LLMs zijn getraind op datasets die **weinig tot geen NLse data bevatten**.

**Europese waardes rondom bias, inclusiviteit en uitlegbaarheid** zijn onvoldoende gegarandeerd in huidige LLMs.

**Digitale soevereinitieit** van Europese taal- en teksttechnologie, geen afhankelijkheid van buitenlandse multinationals

**Privacy en IP-bescherming** van burgers

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.



LEAM:AI

## Große KI-Modelle

FÜR DEUTSCHLAND

Machbarkeitsstudie 2023



### 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?

♦ WSJ NEWS EXCLUSIVE

### Europe to ChatGPT: Disclose Your Sources

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

GPT-NL

# Motivatie

Veel huidige LLMs zijn getraind op datasets die **weinig tot geen NLse data bevatten**.

**Europese waardes rondom bias, inclusiviteit en uitlegbaarheid** zijn onvoldoende gegarandeerd in huidige LLMs.

**Digitale soevereiniteit** van Europese taal- en teksttechnologie, geen afhankelijkheid van buitenlandse multinationals

**Privacy en IP-bescherming** van burgers

GENERATIEVE AI



Overheidsbrede visie  
**Generatieve AI**

**GPT-NL**

# GPT-NL wordt...

Een Nederlands taalmodel, ontworpen voor Nederlandse taal en cultuur, normen, en waarden:

- Open en transparant
- Betrouwbaar en compliant
- Soeverein

# GPT-NL wordt... **open en transparant**



Openheid en transparantie vragen kwetsbaarheid, eerlijkheid, en goede communicatie.

- Amerikaanse en andere aanbieders zijn ondoorzichtig.
- Er kan niet worden gegarandeerd dat het model met Nederlandse/Europese waarden is ontworpen.

# GPT-NL wordt... open en transparant

## 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.

(Publieke) commitment naar  
onze RAI ambities, helpt om  
onszelf accountable te houden.

Publieke waarden  
kunnen wringen

# GPT-NL wordt... **open en transparant**

Openheid en transparantie vragen kwetsbaarheid, eerlijkheid, en goede communicatie.

- Amerikaanse en andere aanbieders zijn ondoorzichtig.
- Er kan niet worden gegarandeerd dat het model met Nederlandse/Europese waarden is ontworpen.

>>> **Open en transparant over de keuzes die tijdens de datacuratie en het trainingsproces worden gemaakt.**

# GPT-NL wordt... **betrouwbaar en compliant**



- Betrouwbaarheid betekent een technisch sterk en rigide model, maar ook dat het volgens de wet is ontworpen. Zo ontstaat er een brug naar **vertrouwen**.

# GPT-NL wordt... betrouwbaar en compliant

## 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   | Coherer 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    |

# GPT-NL wordt... betrouwbaar en compliant

## 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 Intelligence (HAI)

|                            | OpenAI  | cohere         | stability.ai        | ANTHROPIC | Google  | BigScience | Meta    | AI21labs | ALEPH AI |
|----------------------------|---------|----------------|---------------------|-----------|---------|------------|---------|----------|----------|
| 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                     | ○ ○ ○ ○ | ○ ○ ○ ○        | ○ ○ ○ ○             | ○ ○ ○ ○   | ○ ○ ○ ○ | ○ ○ ○ ○    | ○ ○ ○ ○ | ○ ○ ○ ○  | ○ ○ ○ ○  |
| Capabilities & limitations | ● ○ ○ ○ | ● ○ ○ ○        | ● ○ ○ ○             | ● ○ ○ ○   | ● ○ ○ ○ | ● ○ ○ ○    | ● ○ ○ ○ | ● ○ ○ ○  | ● ○ ○ ○  |
| Bias & discrimination      | ● ○ ○ ○ | ● ○ ○ ○        | ● ○ ○ ○             | ● ○ ○ ○   | ● ○ ○ ○ | ● ○ ○ ○    | ● ○ ○ ○ | ● ○ ○ ○  | ● ○ ○ ○  |
| Moderation                 | ● ○ ○ ○ | ● ○ ○ ○        | ● ○ ○ ○             | ● ○ ○ ○   | ● ○ ○ ○ | ● ○ ○ ○    | ● ○ ○ ○ | ● ○ ○ ○  | ● ○ ○ ○  |
| Downstream documentation   | ● ○ ○ ○ | ● ○ ○ ○        | ● ○ ○ ○             | ● ○ ○ ○   | ● ○ ○ ○ | ● ○ ○ ○    | ● ○ ○ ○ | ● ○ ○ ○  | ● ○ ○ ○  |
| Totals                     | 25 / 48 | 23 / 48        | 22 / 48             | 7 / 48    | 27 / 48 | 36 / 48    | 21 / 48 | 8 / 48   | 5 / 48   |

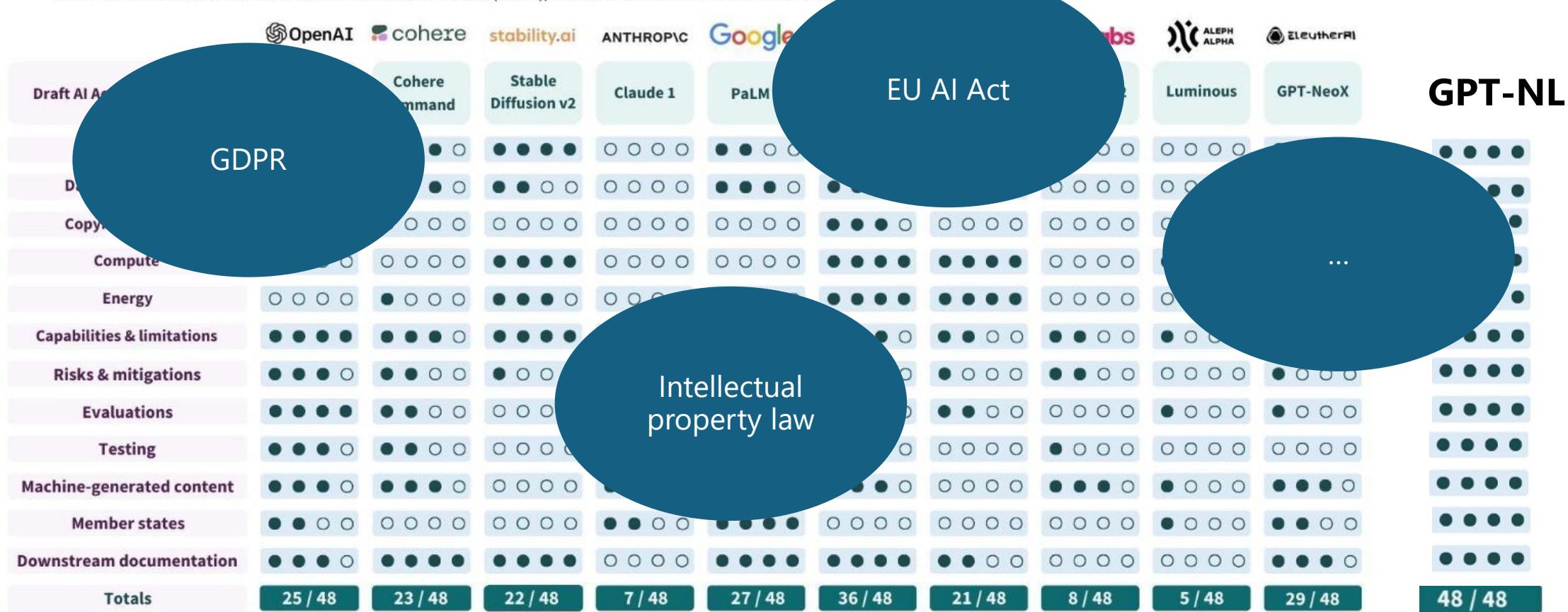
"Impossible": OpenAI admits ChatGPT can't exist  
without pinching copyrighted work

(Klik om link te openen)

# GPT-NL wordt... betrouwbaar en compliant

## 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 Intelligence (IHC), and University of California Berkeley (UCB)



**GPT-NL**

# GPT-NL wordt... **betrouwbaar en compliant**

- Betrouwbaarheid betekent een technisch sterk en rigide model, maar ook dat het volgens de wet is ontworpen. Zo ontstaat er een brug naar **vertrouwen**.

**>>> Omdat we GPT-NL in lijn met de AVG en de AI Act ontwikkelen, worden bronnen vermeden die privacy- of IP-rechten schenden. Zo beschermen we de burger: zowel de gebruiker als het datasubject.**

# GPT-NL wordt... **soeverein**



**Digitale soevereiniteit** betekent

- **controle hebben** over het ontwerp en gebruik van digitale systemen, algoritmen en het verwerken van data.
- Dat geeft ons meer **zeggenschap** over verdere ontwikkeling van technologieën.

>>> **Met GPT-NL vermindert de afhankelijkheid van (Amerikaanse) tech bedrijven voor dataopslag en rekenkracht en nemen we als Europa een sterkere positie in op AI.**

The background of the slide features a subtle, abstract design of radial light streaks. These streaks are composed of thin, glowing lines that radiate from a central point on the left side of the frame, creating a sense of depth and motion. The colors of the streaks transition through a spectrum, including shades of blue, green, yellow, and red, which are partially obscured by a soft, white glow.

# Wat denken jullie?

Of zijn er vragen?

# De grootste uitdaging...

(van Responsible AI tegen "Big Tech" AI)

# We kunnen het niet alleen!

Responsible AI vraagt om samenwerking, eerlijkheid, en open discussie. **Wat zijn jullie ervaringen?**

We staat open voor:

- jullie ideeën voor een sterk, en soeverein AI-ecosysteem binnen Nederland;
- alle hulp voor een rijke, diverse dataset. Alleen samen bouwen we GPT-NL!

# Hartelijk dank!

[info@gpt-nl.nl](mailto:info@gpt-nl.nl)

# GPT-NL.nl

**TNO** innovation  
for life



**SURF**

