

Responsible AI

Amsterdam, 19 September '24

About me



Lieke Dom

- Consultant Responsible Innovation @ TNO Vector
- Fairness and harms of AI
- In GPT-NL responsible for upholding Responsible Innovation practices (and communication)

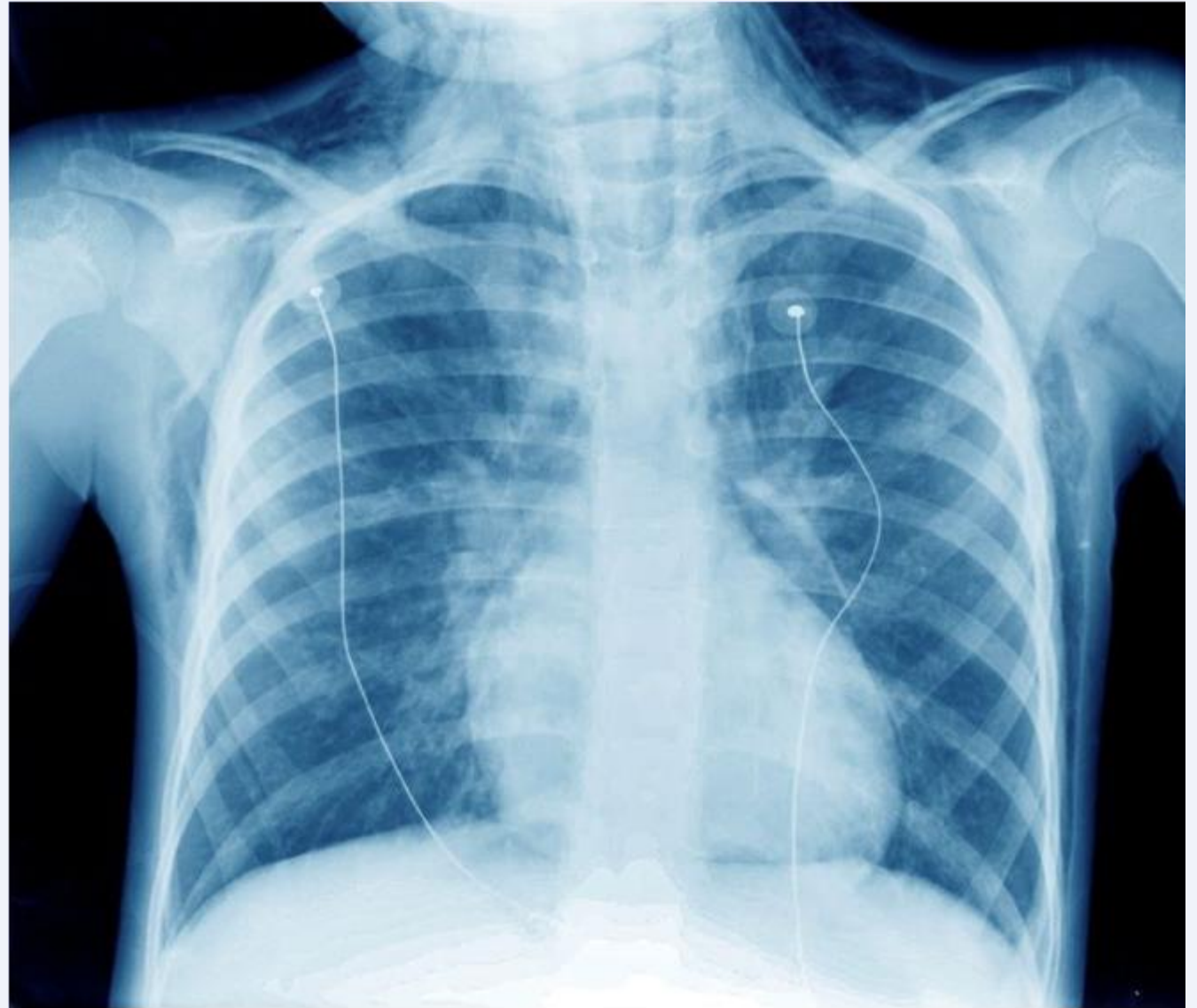
Responsible AI: what does it mean to you?

A Responsible AI dilemma

Researchers in the US discovered something odd: they've been able to train a model that with high accuracy could detect from x-ray imagery if someone was a white or black person. However, the researchers could not track down *what* information from the image was used to distinguish between lighter or darker skin.

The researchers have tried to remove information from the images (e.g. by masking, blurring or filtering) but the model kept on giving accurate results.

What do you think about this?



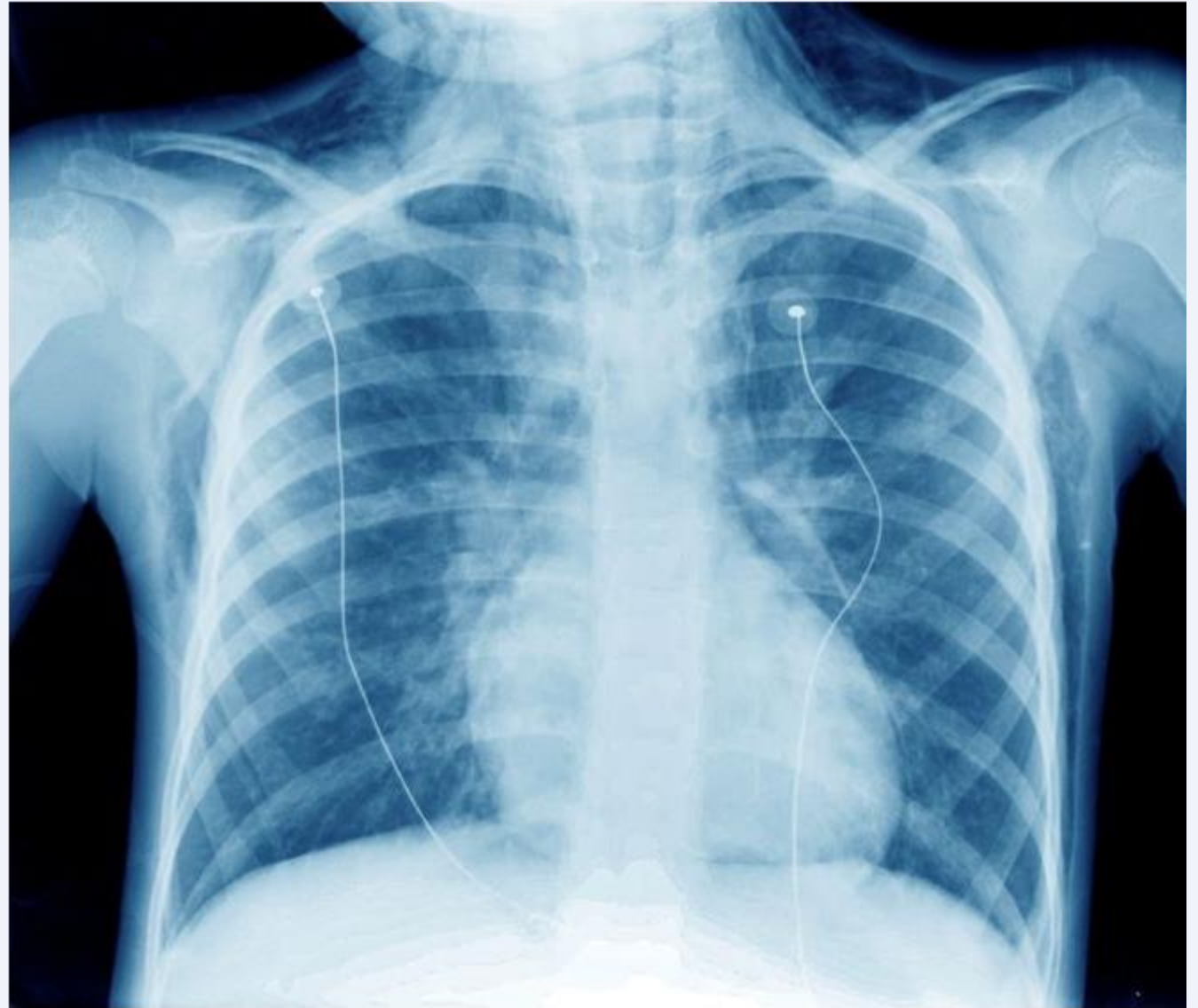
A Responsible AI dilemma

Researchers in the US discovered something odd: they've been able to train a model that with high accuracy could detect from x-ray imagery if someone was a white or black person. However, the researchers could not track down *what* information from the image was used to distinguish between lighter or darker skin.

The researchers have tried to remove information from the images (e.g. by masking, blurring or filtering) but the model kept on giving accurate results.

Fundamental questions.

- What is the impact if a model is using proxies to discriminate without the awareness of developers and end-users?
- Is accuracy the most important result?



GPT-NL



TNO innovation
for life



About us

About us



TNO innovation
for life

SURF



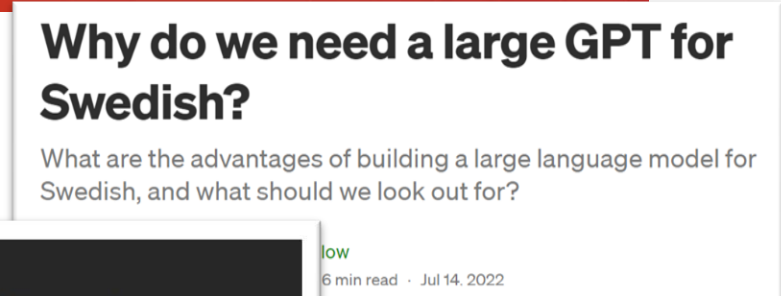
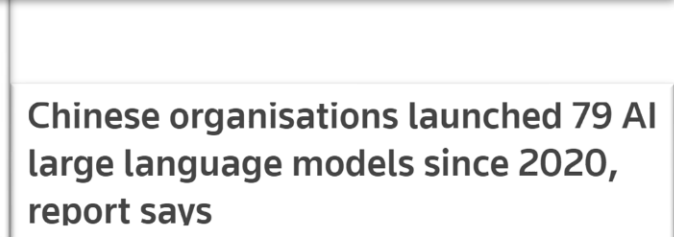
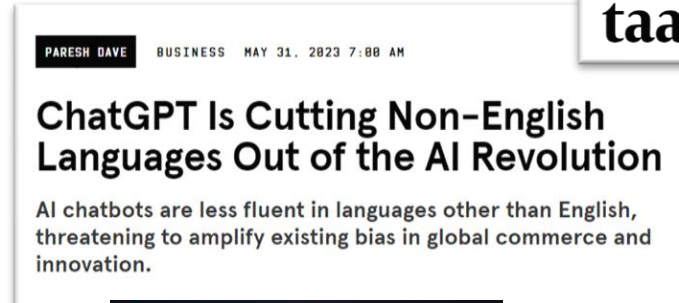
Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid



GPT-NL

Motivation

- In current LLMs, **privacy, data and IP** is not enough protected.
- Current LLMs are trained on datasets that contain **little or no Dutch data**.
- European values with regard to **bias, inclusivity and explainability** are not sufficiently guaranteed in current LLMs because **transparency is lacking**.
- Need for **digital sovereignty** of European language, speech and text technologies, instead of dependence on American multinationals.



Motivation

- In current LLMs, **privacy, data and IP** is not enough protected.
- Current LLMs are trained on datasets that contain **little or no Dutch data**.
- European values with regard to **bias, inclusivity and explainability** are not sufficiently guaranteed in current LLMs because **transparency is lacking**.
- Need for **digital sovereignty** of European language, speech and text technologies, instead of dependence on American multinationals.



Motivation

- In current LLMs, **privacy, data and IP** is not enough protected.
- Current LLMs are trained on datasets that contain **little or no Dutch data**.
- European values with regard to **bias, inclusivity and explainability** are not sufficiently guaranteed in current LLMs because **transparency is lacking**.
- Need for **digital sovereignty** of European language, speech and text technologies, instead of dependence on American multinationals.

GPT-NL is **built from scratch** in accordance with **AI Act, GDPR, and IP law**.

GPT-NL will be trained +/- **50/50 on Dutch and English data**.

GPT-NL will be **transparent, inclusive, and fair** as possible.

Having an LLM from Dutch ground will enforce the Netherland's and Europa's (knowledge) **position on Artificial Intelligence**.

Responsible AI can relate to the technology, the organisation, or the bigger impact.

Examples & dilemmas

- In current LLMs, **privacy, data and IP** is not enough protected.

GPT-NL is **built from scratch** in accordance with **AI Act, GDPR, and IP law**.

- Current LLMs are trained on datasets that contain **little or no Dutch data**.

GPT-NL will be trained +/- **50/50 on Dutch and English data**.

- European values with regard to **bias, inclusivity and explainability** are not sufficiently guaranteed in current LLMs because **transparency is lacking**.

GPT-NL will be **transparent, inclusive, and fair** as possible.

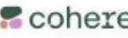






- Need for **digital sovereignty** of European language, speech and text technologies, instead of dependence on American multinationals.

Having an LLM from Dutch ground will enforce the Netherland's and Europa's (knowledge) **position on Artificial Intelligence**.

1: Data, compliance, and quality

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 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

1: Data, compliance, and quality

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 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

1: Data, compliance, and quality

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
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	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○
Privacy	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○
Machinability	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○
Model	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○
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

1: Data, compliance, and quality

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 ALPHA	EleutherAI
Draft AI Act Requirements	GPT-4	Command	Flan	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

“Good enough”

Inspiring others





1: Data, compliance, and quality

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 ALPHA	EleutherAI
Draft AI Act										
Cohere Command										
Stable Diffusion v2										
Claude 1										
PaLM 2										
BLOOM										
LLaMA										
Jurassic-2										
Luminous										
GPT-NeoX										
GDPR										
EU AI Act										
Intellectual property law										
...										
Draft AI Act										
Coherent										
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

Motivation

- In current LLMs, **privacy, data and IP** is not enough protected. 
 - Current LLMs are trained on datasets that contain **little or no Dutch data**. 
 - European values with regard to **bias, inclusivity and explainability** are not sufficiently guaranteed in current LLMs because **transparency is lacking**. 
 - Need for **digital sovereignty** of European language, speech and text technologies, instead of dependence on American multinationals. 
- GPT-NL is **built from scratch** in accordance with **AI Act, GDPR, and IP law**.
- GPT-NL will be trained +/- **50/50 on Dutch and English data**.
- GPT-NL will be **transparent, inclusive, and fair** as possible.
- Having an LLM from Dutch ground will enforce the Netherland's and Europa's (knowledge) **position on Artificial Intelligence**.

2: Transparency

Public commitments help us to be held accountable and ensure auditability.

With regard to transparency, we commit to:

- Publishing code under an open source license.
- Publishing data sheets and model cards for all data sets and models according to industry standards.
- Ambition to publish datasets used to train GPT-NL. Although some datasets will fall under a certain license, therefore, we'll dedicate extra attention on developing transparency mechanisms for these datasets.

GPT-NL Home Contact

Over GPT-NL **Commitments** Samenwerken Planning Veelgestelde vragen

Home / **Commitments**

Onze commitments

Vij werken met toewijding aan de ontwikkeling van een transparant en betrouwbaar taalmodel. Hieronder lees je onze commitments om tot dit resultaat te komen.

Onze commitments voor het bouwen van betrouwbare AI

Wij committeren ons aan het bouwen van een betrouwbaar taalmodel dat in lijn is met de [Ethics Guidelines for Trustworthy AI](#) van de EU waarin staat dat betrouwbare AI-systemen rechtmatig, ethisch en robuust moeten zijn. Onze commitments om een betrouwbaar systeem te bouwen, helpen we binnen aan dit doel te bereiken.

Deze commitments helpen om de ambities van het GPT-NL-project te verduidelijken en om ervoor zorgen dat onze (publieke) belanghebbenden weten wat ze van ons kunnen verwachten. We zijn er ook van overtuigd dat de commitments ons helpen om een weg te vinden in de vele uitdagingen die we tegen zullen komen. Ons opzet is open en onze overwegingen en ontwerpen te delen, publiceren we deze commitments hier. De specifieke implementatie van de commitments zal zich in de loop van de tijd ontwikkelen.

1. Met betrekking tot het proces van het project, committeren wij ons aan:

- het publiceren van dit document met commitments. We zullen de tijd met commitments regelmatig herzien om feedback te verwelken en publiekelijk te rapporteren over eventuele wijzigingen in de commitments.
- het publiceren van een document waarin we de besluitvorming tijdens het opbouwen van onze datasets beschrijven.

2. Met betrekking tot de eindproducten, committeren wij ons aan:

- het publiceren van een succesrapport voor het GPT-NL-project, of twee weken na afsluiting van het moment waarop het project voor ons is gekoppeld. Deze rapportage dient uiterlijk aan het einde van de dataverzamelingsmijstap te worden gepubliceerd.
- het publiceren van een overzicht van de voorbeelden eindproducten van het project, inclusief een beschrijving van het voorbeeld, de open toegankelijkheid en de licentie. Het overzicht dient uiterlijk aan het einde van de dataverzamelingsmijstap te worden gepubliceerd. We streven naar een zo open mogelijk eindproduct, maar omdat dit afhankelijk is van overeenkomsten met dataleveranciers kunnen we dit nog niet garanderen.
- Een duidelijk beschreven en permissieve licentie model bij elk eindproduct.

3. Met betrekking tot transparantie, committeren wij ons aan:

- het openbaar publiceren van de code onder een open source-licentie.
- het publiceren van data sheets en modelcards voor alle datasets en modellen (eindproducten) volgens best practices uit de industrie.
- de ambities om de gebruikte datasets voor het trainen van GPT-NL standaard vrij te geven en te publiceren. Sommige datasets kunnen echter onder een licentie vallen, waardoor volledige publicatie wordt beperkt. Voor die datasets zullen we expliciet aandacht besteden aan het creëren van andere transparantiemechanismen.

4. Met betrekking tot ons datagebruik, committeren wij ons aan:

- We garanderen alleen content voor het trainen van GPT-NL als de dataleverancier de juiste rechten heeft om ons hiervoor een licentie te verschaffen. Dit betekent dat de dataleverancier ofwel de eigenaar moet zijn van de auteursrechten of dataleverancier in de dataset, ofwel geldige licentierechten heeft gekregen van de desbetreffende eigenaar.
- We trainen GPT-NL niet op informatie die onderworpen is aan wetgeving of contractuele vertrouwelijkheidsovereenkomsten (zoals vertrouwelijke patiëntinformatie of bedrijfsgegevens).
- We richten ons specifiek op het detecteren, filteren en verwijderen van persoonlijke informatie uit de trainingsgegevens.
- We richten ons specifiek op het detecteren, filteren en verwijderen van schadelijke inhoud - zoals geweldige, criminele of discriminerende inhoud of haatdragend taalgebruik - uit onze trainingsgegevens.

5. Met betrekking tot diversiteit en inclusie, committeren wij ons aan:

- Om vooroordelen in het model zo goed mogelijk te beperken, creëren we een basisdataset die zoveel mogelijk groepen vertegenwoordigt.
- We betrekken ondervertegenwoordigde groepen bij het verbeteren van het model in de finetuning-fase.

6. Met betrekking tot onze belanghebbenden en de communicatie richting het publiek, committeren wij ons aan:

- We publiceren ons communicatieplan en zullen elk kwartaal (om de drie maanden) een update communicatie richting het publiek.
- We publiceren regelmatig (openbare) rapporten over bevindingen die binnen het project zijn gekomen, inclusief rapportages over juridische en ethische dilemma's en bevindingen.
- We rapporteren over de conclusies uit overleg met stakeholders (zie onder):
- Overleg met stakeholders wordt in ieder geval georganiseerd voor:
 - betrouwbare bij de voorbereiding van de finetuning-fase.
 - raadpleging over methoden om de prestaties van het model te verbeteren (op technische en maatschappelijke benchmarks).
- Raadplegingen van belanghebbenden worden publiekelijk aangekondigd op de website en sociale media van GPT-NL.

THOUGHT | |

Cookie Privacy statement Disclaimer Toegankelijkheid GitHub Blogging Face

in

2: Transparency

Public commitments help us to be held accountable and ensure auditability.

With regard to transparency, we commit to:

- Publishing code under an open source license.
- Publishing data sheets and model cards for all data sets and models according to industry standards.
- **Ambition to publish datasets used to train GPT-NL. Although some datasets will fall under a certain license, therefore, we'll dedicate extra attention on developing transparency mechanisms for these datasets.**

Public values can conflict, especially when it comes to privacy.

GPT-NL Home Contact

Over GPT-NL **Commitments** Samenwerken Planning Veelgestelde vragen

Home / **Commitments**

Onze commitments

Vij werken met toewijding aan de ontwikkeling van een transparant en betrouwbaar taalmodel. Hieronder lees je onze commitments om tot dit resultaat te komen.

Onze commitments voor het bouwen van betrouwbare AI

Wij committeren ons aan het bouwen van een betrouwbaar taalmodel dat in lijn is met de [Ethics Guidelines for Trustworthy AI](#) van de EU waarin staat dat betrouwbare AI-systemen rechtmatig, ethisch en robuust moeten zijn. Onze commitments om een betrouwbaar systeem te bouwen, hebben we binnen zes themas gebundeld.

Deze commitments helpen om de ambities van het GPT-NL-project te verduidelijken en om ervoor zorgen dat onze (publieke) belanghebbenden weten wat ze van ons kunnen verwachten. We zijn er ook van overtuigd dat de commitments ons helpen om een weg te vinden in de vele uitdagingen die we tegen zullen komen. Ons opzet is open over onze overwegingen en onze besloten te delen, publiceren we deze commitments hier. De specifieke implementatie van de commitments zal zich in de loop van de tijd ontwikkelen.

1. Met betrekking tot het proces van het project, committeren wij ons aan:

- het publiceren van dit document met commitments. We zullen de tijd met commitments regelmatig herzien om feedback te verwelken en publiekelijk te rapporteren over eventuele wijzigingen in de commitments.
- het publiceren van een document waarin we de besluitvorming tijdens het opbouwen van onze datasets beschrijven.

2. Met betrekking tot de eindproducten, committeren wij ons aan:

- het publiceren van een succesverhaal voor het GPT-NL-project, ofwel een omschrijving van het moment waarop het project voor ons is geklaard. Deze definitie dient uiterlijk aan het einde van de dataverzameling-mijstap te worden gepubliceerd.
- het publiceren van een overzicht van de beoogde eindproducten van het project, inclusief een beschrijving van het beoogde doel, de open toegankelijkheid en de licentie. Het overzicht dient uiterlijk aan het einde van de dataverzameling-mijstap te worden gepubliceerd. We streven naar een zo open mogelijk eindproduct, maar omdat dit afhankelijk is van overeenkomsten met dataleveranciers kunnen we dit nog niet garanderen.
- Een duidelijk beschreven en permissieve licentie model bij elk eindproduct.

3. Met betrekking tot transparantie, committeren wij ons aan:

- het openbaar publiceren van de code onder een opensource-licentie.
- het publiceren van datasheets en modelcards voor alle datasets en modellen (eindproducten) volgens best practices uit de industrie.
- de ambities om de gebruikte datasets voor het trainen van GPT-NL standaard vrij te geven en te publiceren. Sommige datasets kunnen echter onder een licentie vallen, waardoor volledige publicatie wordt beperkt. Voor die datasets zullen we expliciet aandacht besteden aan het creëren van andere transparantiemechanismen.

4. Met betrekking tot ons datagebruik, committeren wij ons aan:

- We zullen alleen content voor het trainen van GPT-NL, als de dataleverancier de juiste rechten heeft, om ons hiervoor een licentie te verschaffen. Dit betekent dat de dataleverancier ofwel de eigenaar moet zijn van de auteursrechten of databevoegten in de dataset, ofwel geldige licentierechten heeft gelvegen van de dataleider/ eigenaar.
- We trainen GPT-NL niet op informatie die onderhevig is aan wetgeving of contractuele vertrouwelijkheidsovereenkomsten (zoals vertrouwelijke patiëntinformatie of bedrijfsgegevens).
- We richten ons specifiek op het detecteren, filteren en verwijderen van persoonlijke informatie uit de trainingsgegevens.
- We richten ons specifiek op het detecteren, filteren en verwijderen van schadelijke inhoud - zoals geweldige, criminele of discriminerende inhoud of haatdragend taalgebruik - uit onze trainingsgegevens.

5. Met betrekking tot diversiteit en inclusie, committeren wij ons aan:

- Om vooroordelen in het model zo goed mogelijk te beperken, creëren we een basisset met de zoveel mogelijk groepen vertegenwoordigt.
- We betrekken ondervertegenwoordigde groepen bij het verbeteren van het model in de fracturing-fase.

6. Met betrekking tot onze belanghebbenden en de communicatie richting het publiek, committeren wij ons aan:

- We publiceren ons communicatie plan en zullen elk kwartaal (om de drie maanden) een update communicatie richting het publiek.
- We publiceren regelmatige (openbare) rapporten over beleidsingen die binnen het project zijn genomen, inclusief rapportages over juridische en ethische dilemma's en beleidsingen.
- We rapporteren over de conclusies uit overleg met stakeholders (zie onder):
- Overleg met stakeholders wordt in ieder geval georganiseerd voor:
 - betrokkenheid bij de voorbereiding van de fracturing-fase.
 - raadpleging over methoden om de prestaties van het model te verbeteren (op technische en maatschappelijke benchmarks).
- Raadplegingen van belanghebbenden worden publiekelijk aangekondigd op de website en sociale media van GPT-NL.

THOUCS | |

Cookie Privacy statement Disclaimer Toegankelijkheid GitHub Blogging Face

in

What would you decide?

Openly publishing anonymised data sets

Minimal privacy risk, but big impact if something happens.

Publishing meta data about the data sets

Therefore limiting public auditability and opportunities for researchers

What would you decide?

Anonymising public figures?

To ensure highest anonymisation as possible?

Not anonymising public figures?

To ensure highest knowledge of Dutch culture, history and persons?

3: Bias?

- We often think about discrimination based on ethnicity, gender, socio-economic aspects, e.t.c.
- What kind of biases can we think of in a language model for health care or medical purposes?

3: Bias?

- We often think about discrimination based on ethnicity, gender, socio-economic aspects, e.t.c.
- What kind of biases can we think of in a language model for health care or medical purposes?

Example 1

- American drug ads vs European regulation on ads



Lady
with a Lamp
(1946 Version)

• The pages of medical history during the last century glow with the names of great women. Florence Nightingale, the "lady with the lamp"... Elizabeth Blackwell, the first American woman to be given the proud degree M.D. ... Drs. Mary Putnam Jacobi... Jane Viola Meyers... Anna Broomall... the list is long. And brilliant. In America today, thanks to the intrepid spirit of these pioneers, 7,250 women doctors carry the lamps they lighted ever further along the path of human service.

According to a recent Nationwide survey: **MORE DOCTORS SMOKE CAMELS THAN ANY OTHER CIGARETTE**

• Men and women in every branch of medicine—113,597 in all—were queried in this nationwide study of cigarette preference. Three leading research organizations made the survey. The gist of the query was—What cigarette do you smoke, Doctor?
The brand named most was Camel!
The rich, full flavor and cool mildness of Camel's superb blend of costlier tobaccos seem to have won the same favor in medical circles as with millions of smokers the world around. If you are a Camel smoker, this preference among doctors will hardly surprise you. If you're not—well, try Camels now.



TRY CAMELS ON YOUR "T-ZONE"
That's T for Taste and T for Throat...the most critical "laboratory" for any cigarette. See how your taste responds to the rich, full flavor of Camel's costlier tobaccos. See how your throat reacts to Camel's cool mildness. On the basis of the experience of many millions of smokers, we believe Camels will suit your "T-Zone" to a "T."
© J. R. Rorick Tobacco Co. Winston-Salem, N. C.

CAMELS Costlier Tobaccos

3: Bias?

- We often think about discrimination based on ethnicity, gender, socio-economic aspects, e.t.c.
- What kind of biases can we think of in a language model for health care or medical purposes?

Example 1

- American drug ads vs European regulation on ads

Example 2

- Difference between The Netherlands and our neighbour Belgium

Verschillen in zorg op de HAP tussen België en Nederland

Door Rob van Kimmenaede

Gepubliceerd 12 augustus 2019 Leestijd 1 minuut

Het kan leerzaam zijn om het gebruik van de HAP tussen landen te vergelijken. Daarom bekeken onderzoekers de verschillen tussen Nederland en België. Zij constateren als belangrijkste verschil: het aantal consulten op de HAP per 1000 inwoners ligt in Nederland 2,3 keer hoger dan in België en na correctie voor telefonische consulten (niet toegestaan in België) is dit verschil nog altijd een factor 1,4.

[Artikelinfo](#)

[Afdrukken](#)

[Delen](#)



Thank you!

GPT-NL

TNO innovation
for life



Let's keep in touch!

Contact:

info@gpt-nl.nl

www.gpt-nl.nl

www.linkedin.com/company/gpt-nl/