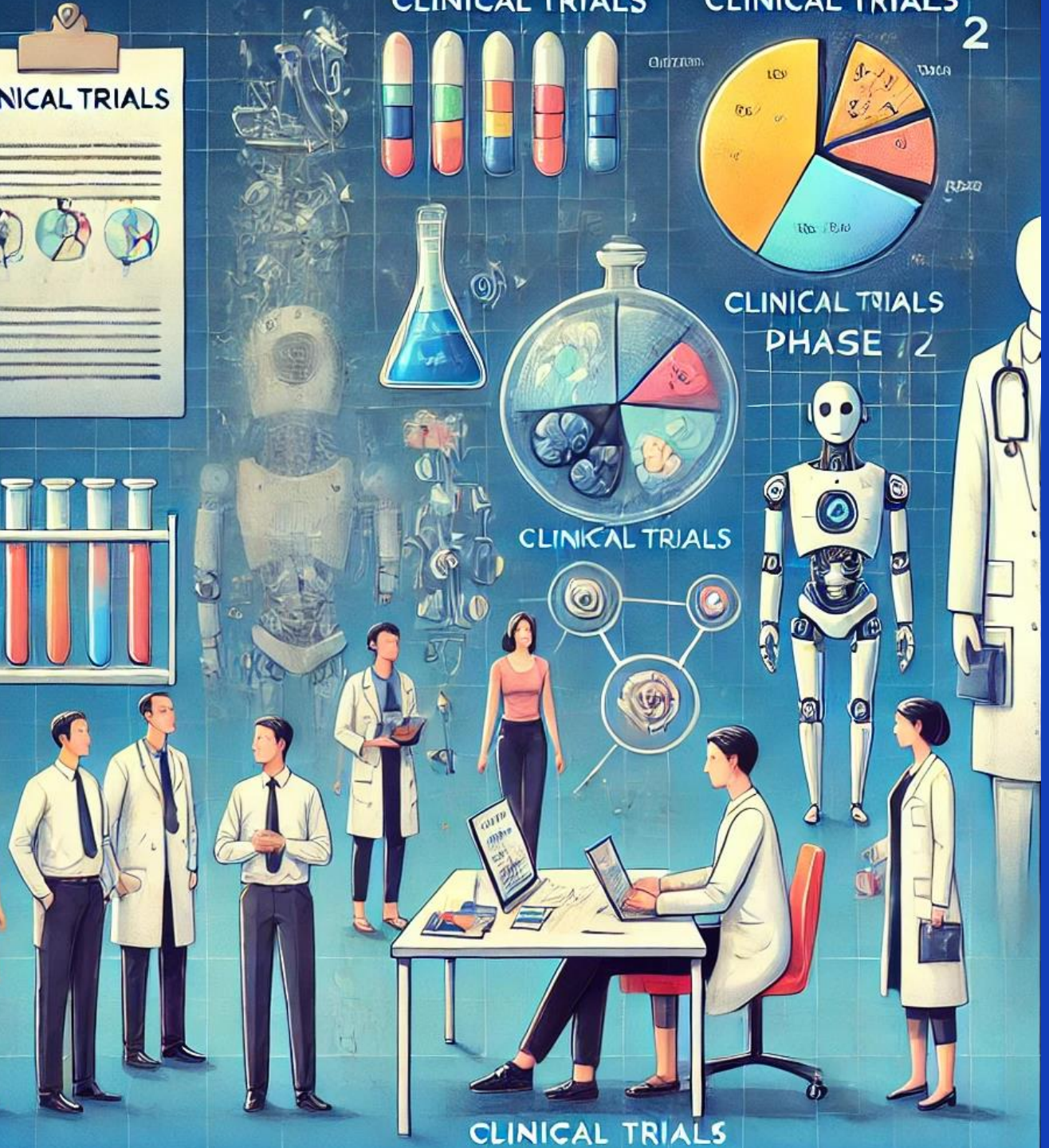
GPT-NL

Towards a public interest Large Language Model

Lieke Dom & Saskia Lensink



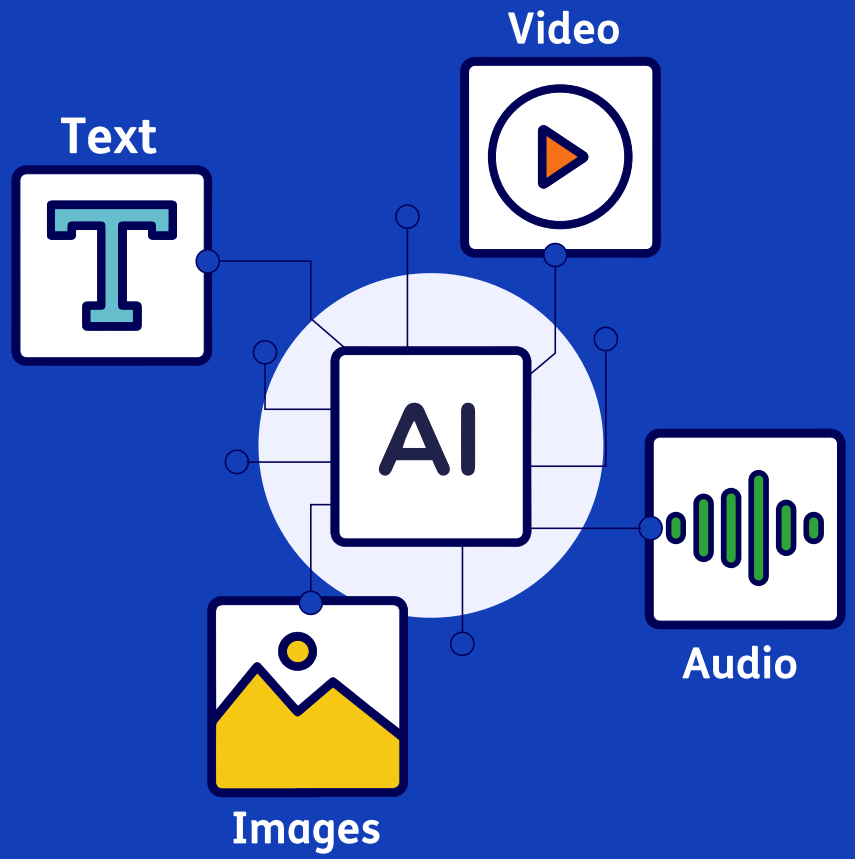
Why do clinical trials take so much time?

Very well regulated because of the potential of a (hugely) negative impact

… how about generative AI?

What is generative AI?

Aa category of AI systems designed to generate new content, such as text, images, audio, and video, by learning patterns from existing data.



ChatGPT

Copilot

udio

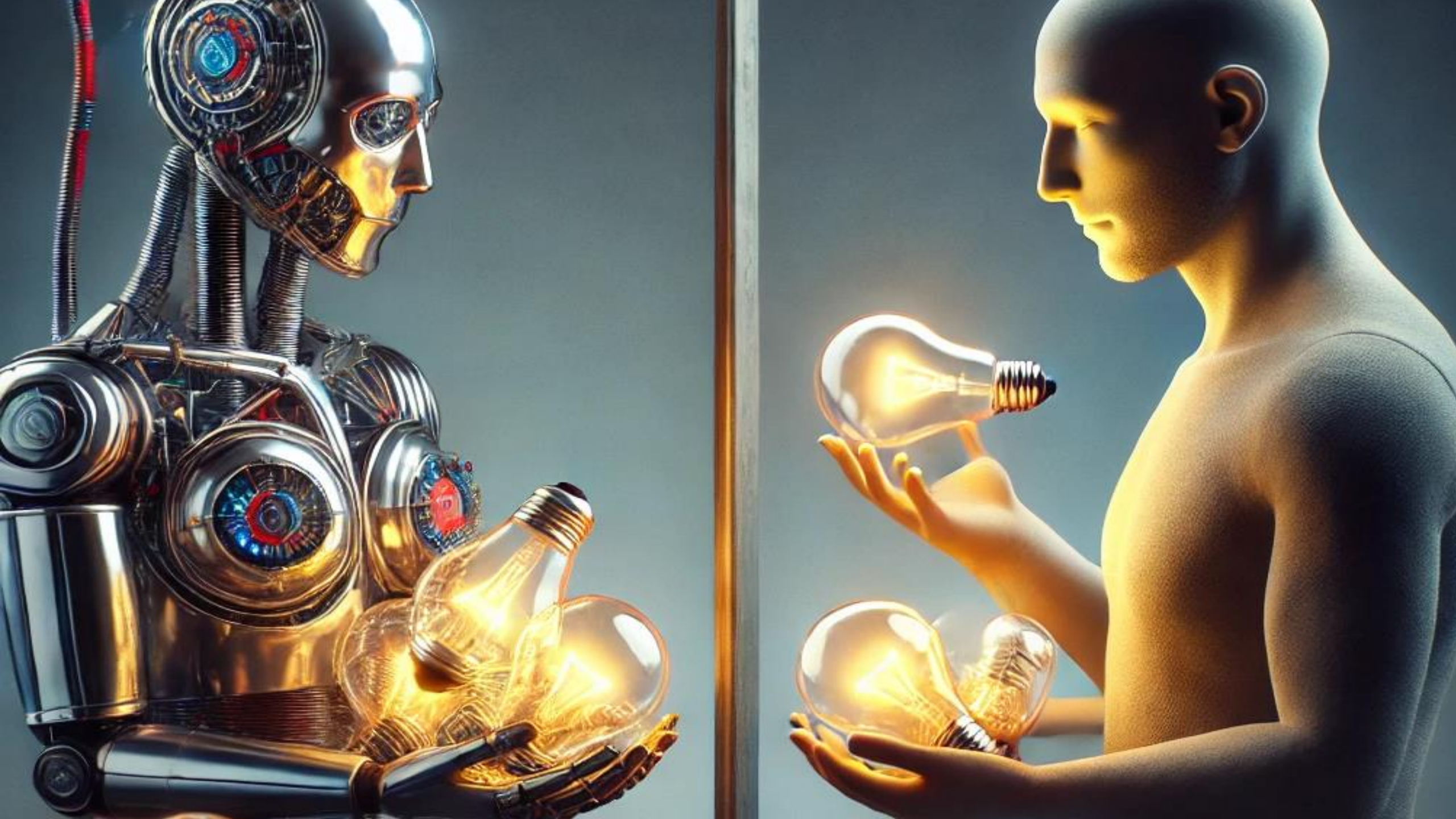
Google AI

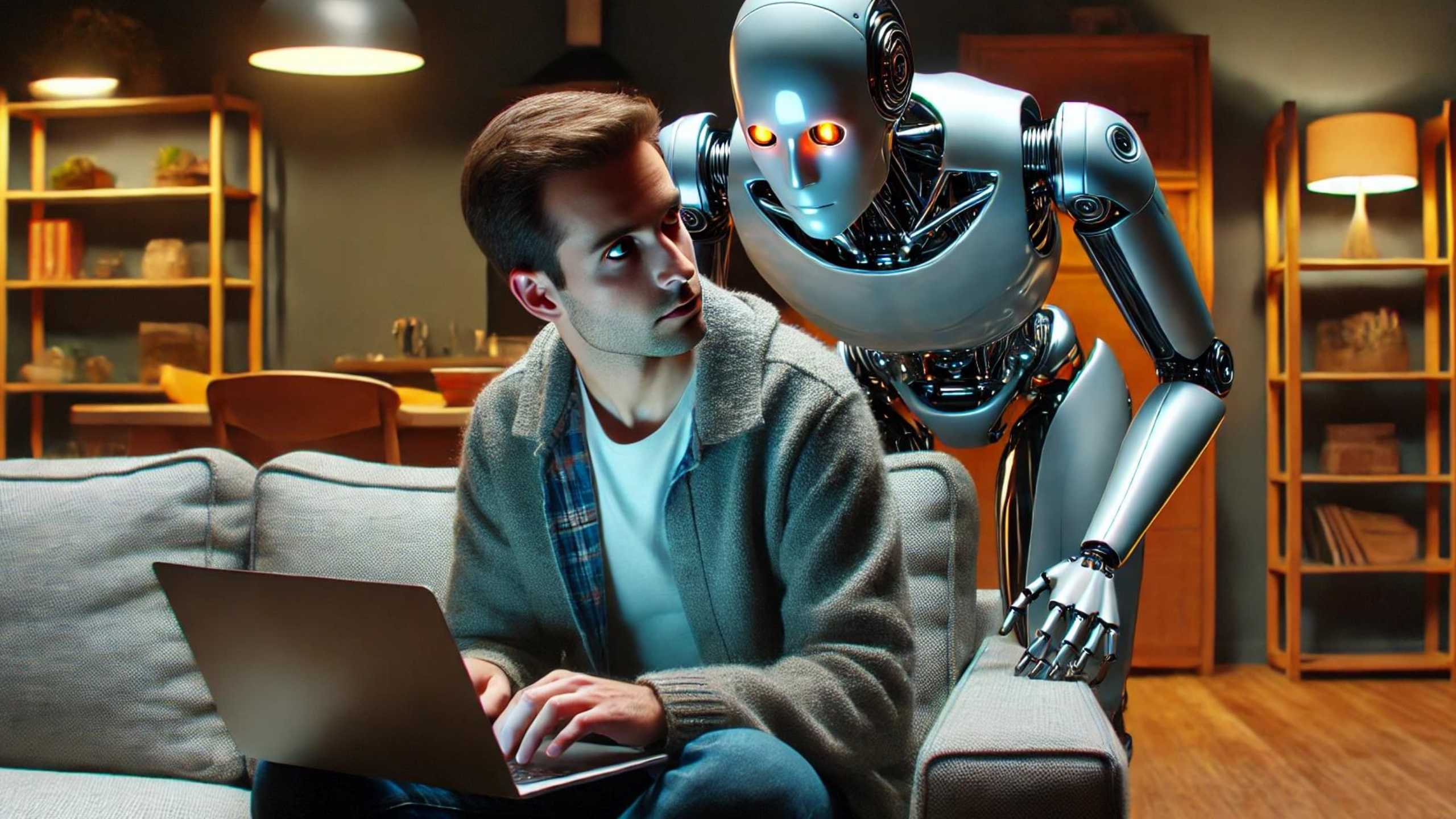
DALL·E

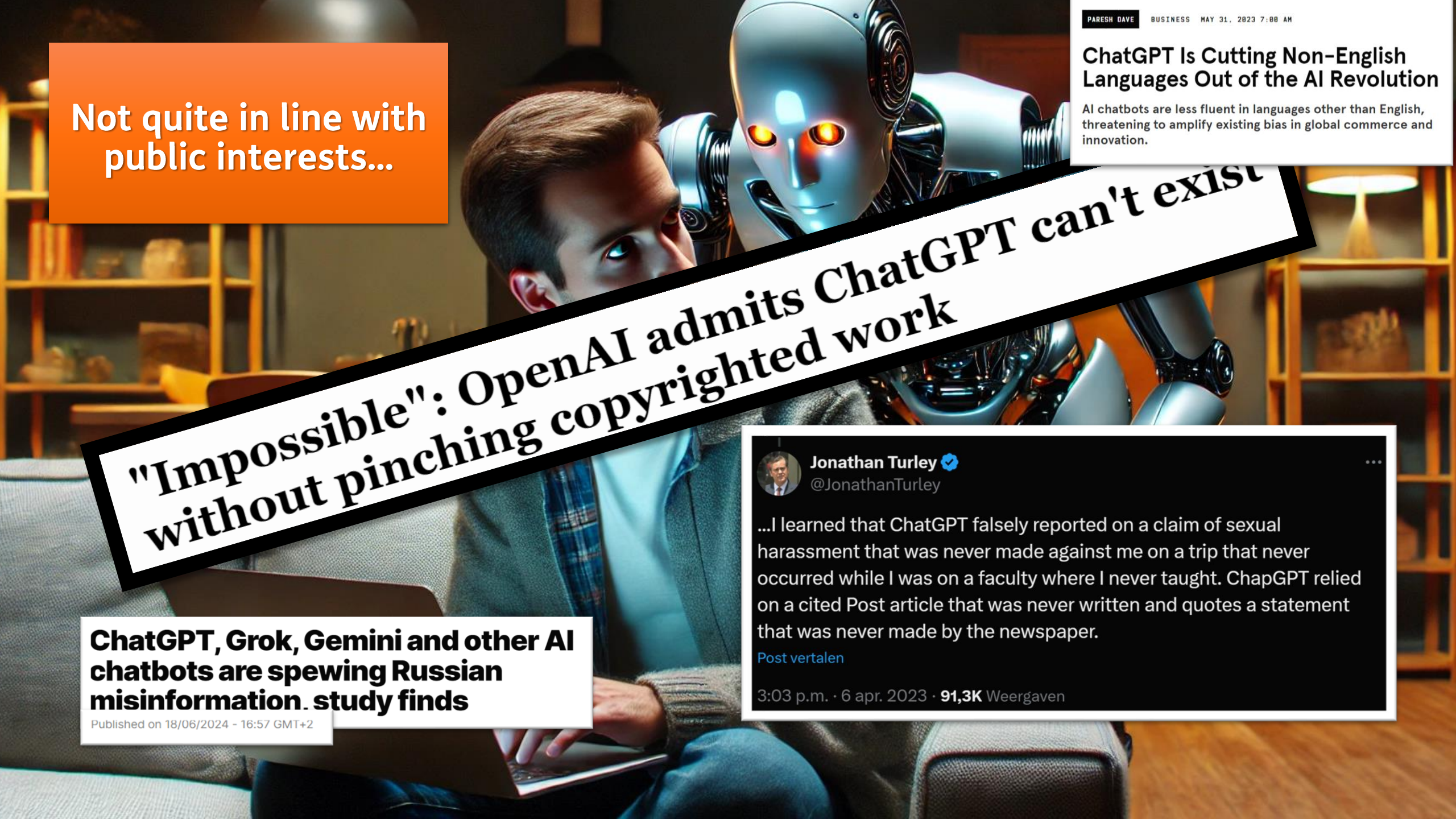
Gemini

GitHub Copilot

Suno







Not quite in line with public interests...

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.

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

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

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

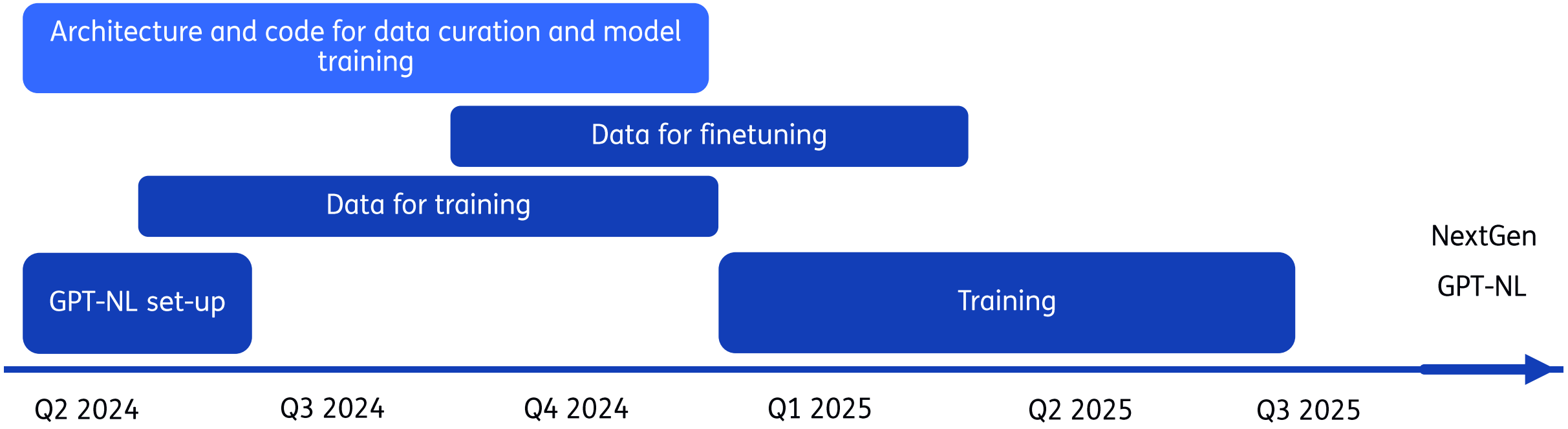
Published on 18/06/2024 - 16:57 GMT+2

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**

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

Planning


Architecture and code for data curation and model training
Data for finetuning
Data for training
GPT-NL set-up
Training
NextGen GPT-NL
Q2 2024
Q3 2024
Q4 2024
Q1 2025
Q2 2025
Q3 2025

EU High-level expert group: ethical guidelines for Trustworthy AI

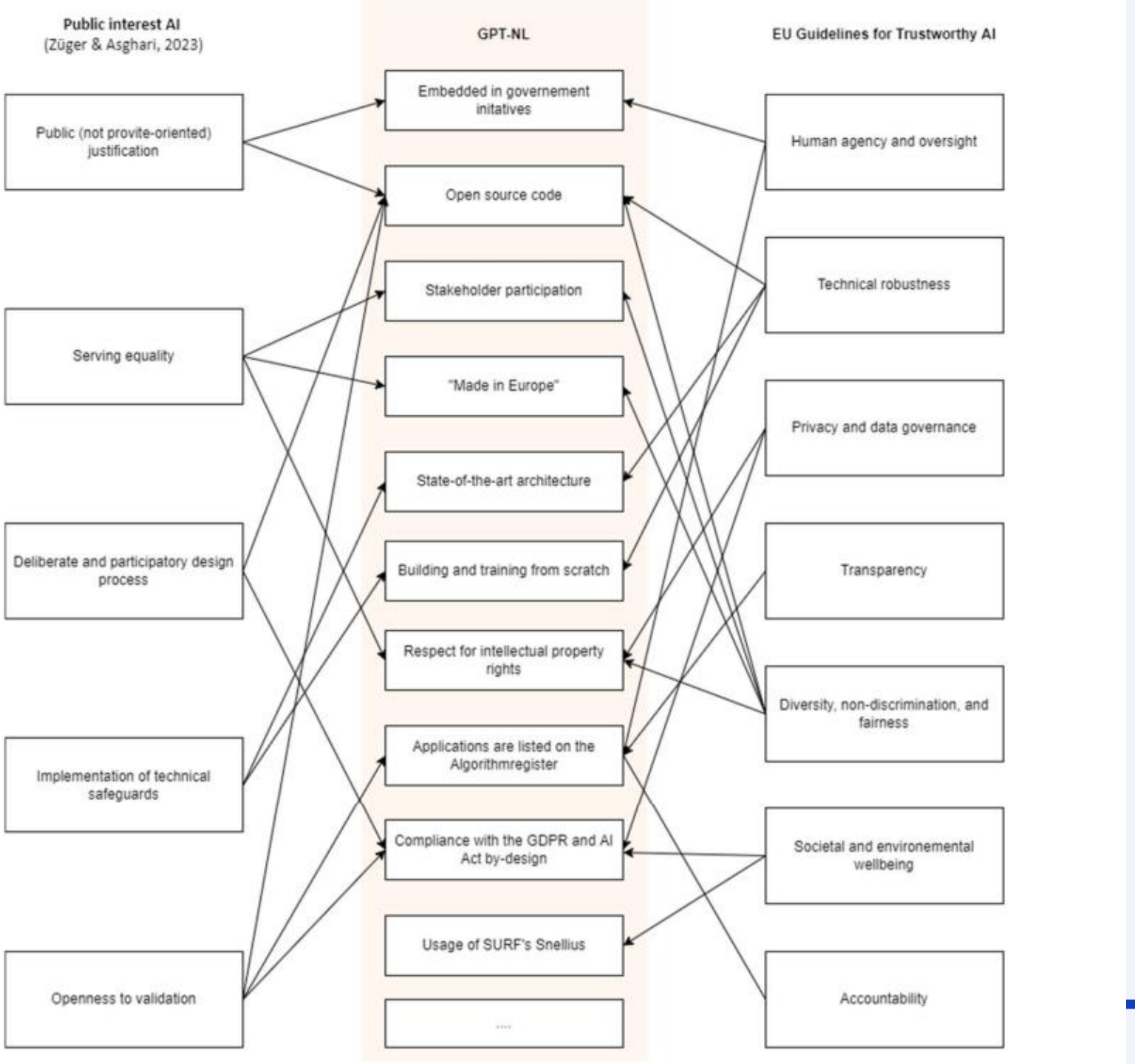
To realize and align GPT-NL with public interests, the model is built in line with the European Union's Guidelines for Trustworthy AI

EU High-level expert group: ethical guidelines for Trustworthy AI

1. **Human agency and oversight**: Including fundamental rights, human agency and human oversight
2. **Technical robustness and safety**: Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility
3. **Privacy and data governance**: Including respect for privacy, quality and integrity of data, and access to data
4. **Transparency**: Including traceability, explainability and communication
5. **Diversity, non-discrimination and fairness**: Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation
6. **Societal and environmental wellbeing**: Including sustainability and environmental friendliness, social impact, society and democracy
7. **Accountability**: Including auditability, minimisation and reporting of negative impact, trade-offs and redress

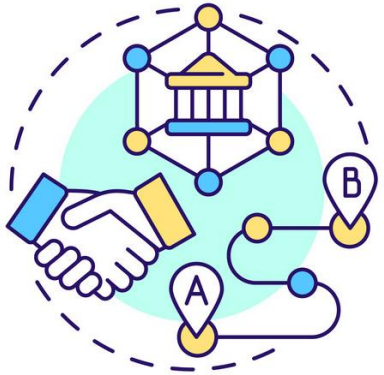
Public interest in political theory Züger & Asghari (2023)

- Alternative lens to the EU's Guidelines
- Five requirements for an AI system to serve the public interest
  - Public (not profit-oriented) justification;
  - Serve equality;
  - Require a deliberation / co-design process;
  - Follow key technical standards & safeguards;
  - Openness to validation.


Public interest AI
(Züger & Asghari, 2023)
GPT-NL
EU Guidelines for Trustworthy AI
Public (not provite-oriented) justification
Serving equality
Deliberate and participatory design process
Implementation of technical safeguards
Openness to validation
Embedded in governement initatives
Open source code
Stakeholder participation
"Made in Europe"
State-of-the-art architecture
Building and training from scratch
Respect for intellectual property rights
Applications are listed on the Algorithmregister
Compliance with the GDPR and AI Act by-design
Usage of SURF's Snellius
....
Human agency and oversight
Technical robustness
Privacy and data governance
Transparency
Diversity, non-discrimination, and fairness
Societal and environemental wellbeing
Accountability

Public (not profit-oriented) justification

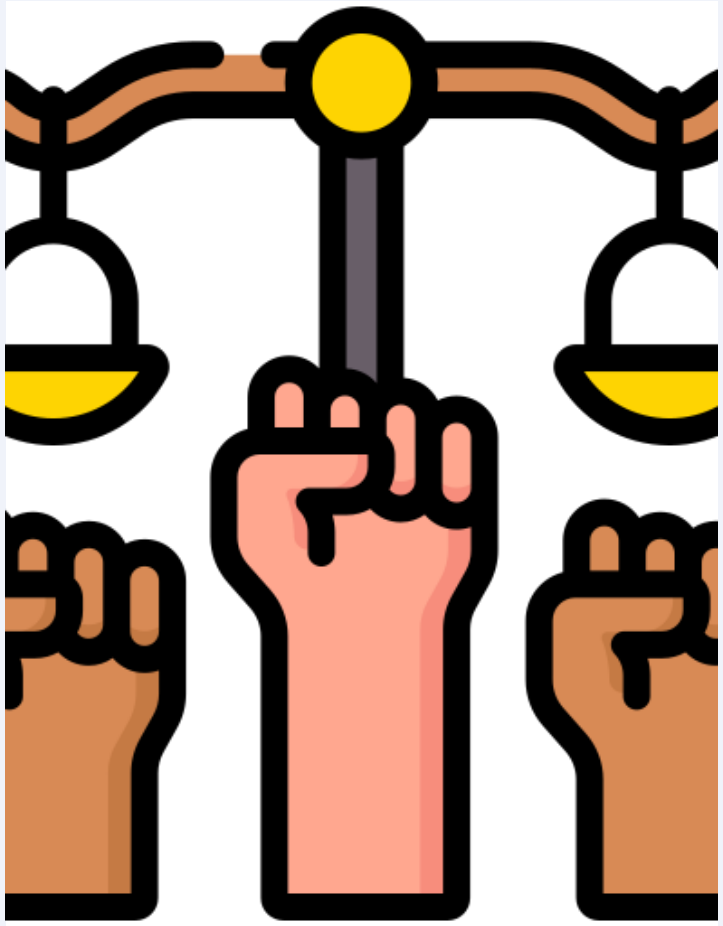
- The Dutch government restricts using American and Chinese LLMs in public organizations due to non-compliance with EU laws.
- GPT-NL will be built from scratch such that compliance and transparency can be realized in every step of the development process.
- GPT-NL aims to offer a compliant, transparent alternative.
- This allows public organizations to innovate with AI - in line with public interest
- Reducing reliance on foreign systems and strengthening European AI expertise (digital sovereignty)



Government Initiatives

Serving Equality and Human Rights

- GPT-NL promotes equality by ensuring equal access to AI technology and upholding human rights through alignment with European regulations.
- The model will be cost-effective, with non-profit licensing based on usage, and trained on representative data from Dutch municipalities, ensuring fairness and reciprocity.



Deliberative and participatory design process

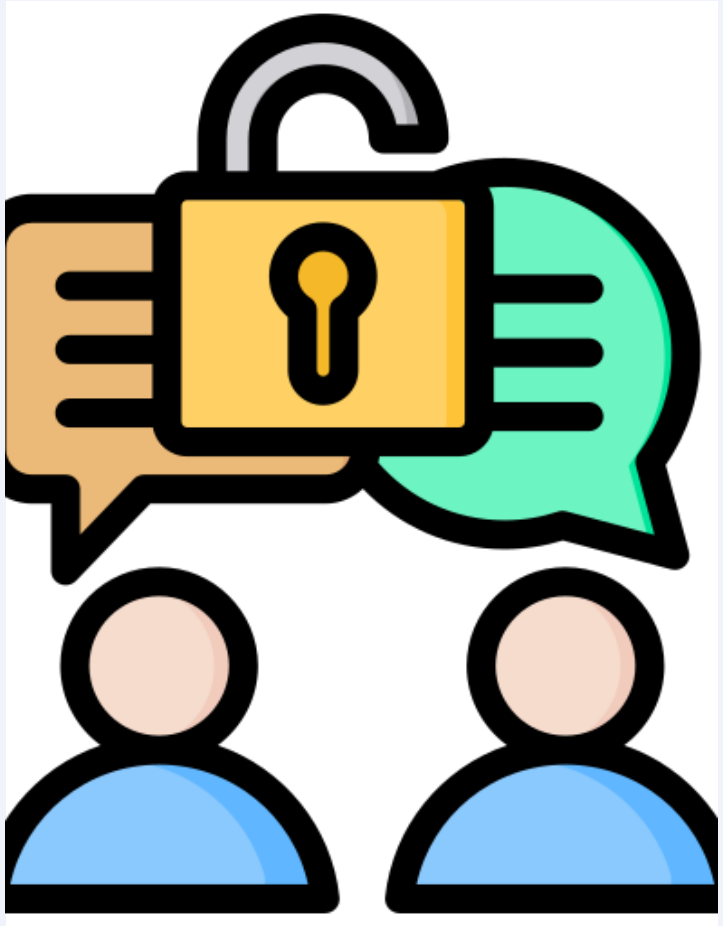
- **Consortium collaboration:** GPT-NL is developed by TNO, SURF, and NFI, ensuring balanced roles. TNO specializes in privacy, anonymization, and societal impact, SURF offers access to the national supercomputer (Snellius) and technical expertise, and NFI provides use cases for testing the model.
- **Participatory data provision:** Organizations can voluntarily share data with GPT-NL, supported by anonymization tools to ensure no sensitive data is shared. GPT-NL also helps data providers develop future use cases.
- **Recent update**: Two stakeholder sessions on (representation) bias were held with 7 experts from various institutions to address representational bias in the development process.

Implementation of technical safeguards

- **Custom model architecture:** GPT-NL is built from scratch, ensuring high-quality data, system accuracy, and data privacy safeguards.
- **Training on Dutch-specific data**: Unlike American LLMs, GPT-NL is trained on high-quality Dutch and English data from consenting providers, enhancing cultural and linguistic accuracy.
- **Compliance with regulations**: GPT-NL development adheres to GDPR, the AI Act, and intellectual property law, ensuring the highest technical standards are applied.

Openness to validation

- **Transparency commitment**: GPT-NL prioritizes transparency by publishing data sheets with performance data and technical specifications, while protecting sensitive information. Development decisions are also shared with regular reporting.
- **Algorithm Register inclusion**: GPT-NL will be listed in the Dutch Algorithm Register to ensure accountability when used by governmental organizations, aligning with upcoming legal requirements for algorithm disclosure by 2024.

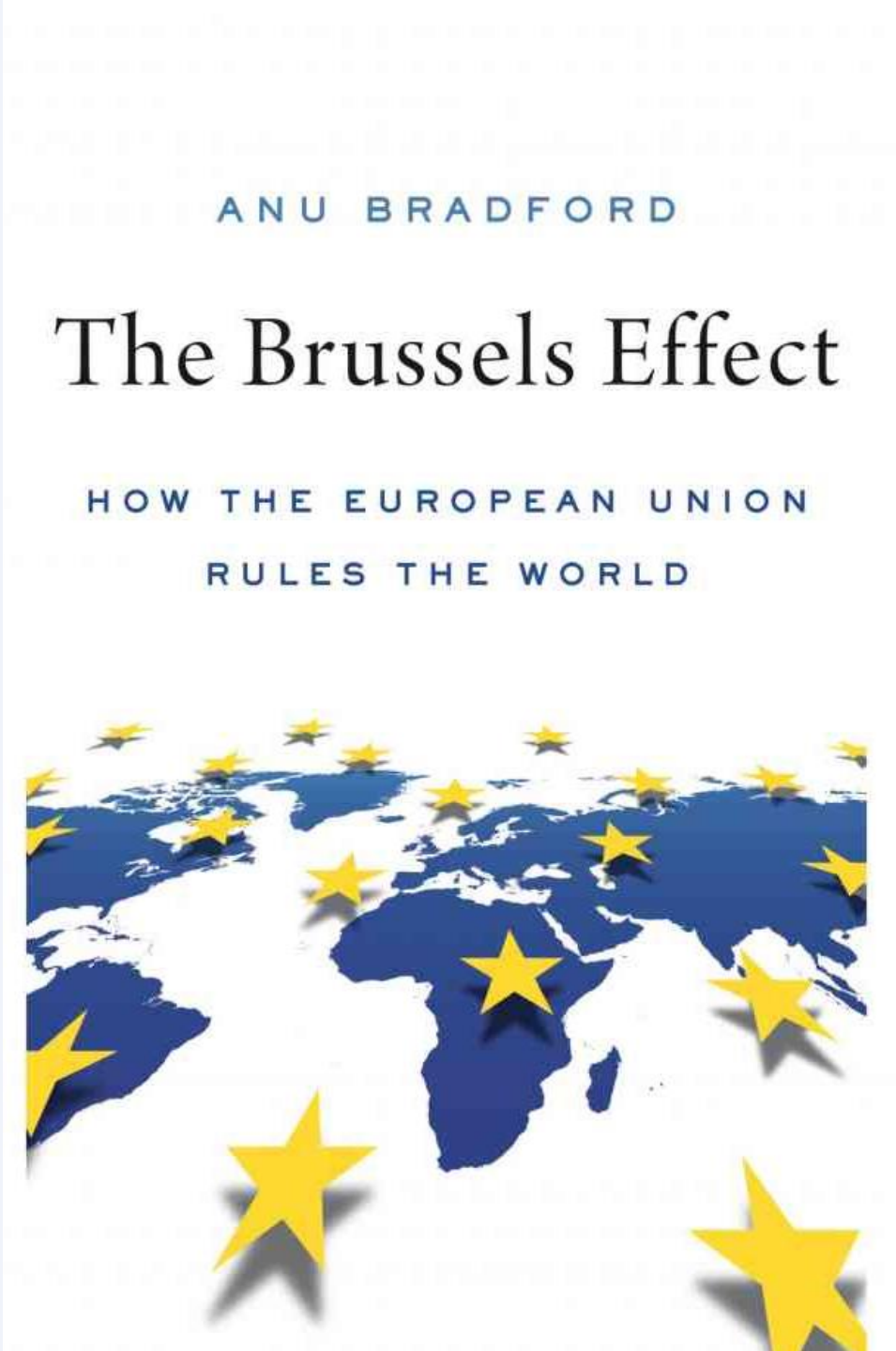


Public interest requirements – self-evaluation

Feature	Score
Public justification	●●●●
Serve equality	●●
Deliberation / co-design process	●●
Follow key technical standards	●●●●
Open for validation	●●●

Discussion

- GPT-NL is a concrete case for the operationalizations of public interest AI guidelines;
- By sharing knowledge and building an ecosystem around GPT-NL, the knowledge position of the Netherlands will increase;
- GPT-NL aims to share knowledge across the European Union, therefore supporting Europe's sovereignty at large
- Hopefully, GPT-NL might cause another 'Brussels effect' showing that it is possible to develop genAI without compromising on lawfulness and ethics



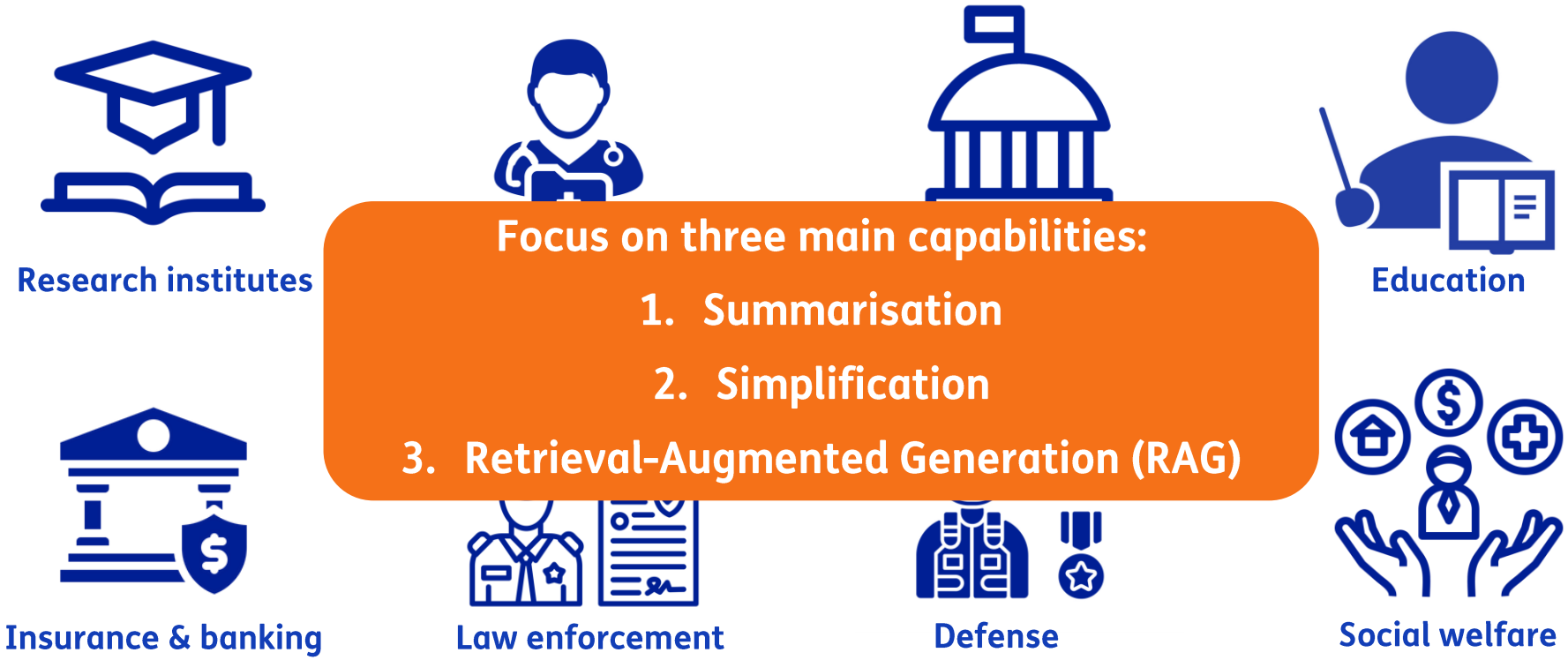
Thank you very much for your attention!

Want to know more? Don’t hesitate to reach out to

lieke.dom@tno.nl

saskia.Lensink@tno.nl

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