GPT-NL

The first year

March 4th 2025 | Saskia Lensink

Why our own LLM?

- In current LLMs, **privacy, data and IP** is not enough protected.
- 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 AI technologies
- Need for a **sustainable and fair data ecosystem**

▲ De Nederlandse driekleur wappert: een eigen AI-taalmodel is in de maak. © Beeldredactie

DeepSeek? Nederland bouwt stilletjes aan eigen betrouwbaar AI-taalmodel: ‘Sentiment kantelt’

De AI-revolutie dendert door met DeepSeek, ChatGPT en CoPilot. Wat weinig mensen weten: achter de schermen wordt gewerkt aan GPT-NL, een door de overheid gesteund, volledig Nederlands AI-taalmodel. Een verhaal over een wildwestmarkt, de kunst van het netjes blijven en waarom Nederland een eigen chatbot wil. „We móéten zelf iets bouwen.”

Joost Dijkgraaf 07-02-25, 10:30 Laatste update: 07-02-25, 12:41

It is essential to restore the security of supply chains for critical technologies by strengthening the EU's capabilities and assets across the entire value chain in terms of end products and service platforms. Moreover, the 'data value loss' (i.e. the amount of EU data transferred to third countries) is today estimated at 90%,[iii] with a long-term risk of loss of industrial know-how. This issue needs to be addressed, especially in light of the crucial role of data in digital developments.

European competitiveness

Draghi: 3 main action areas

(…)

3. increasing security & reducing dependencies

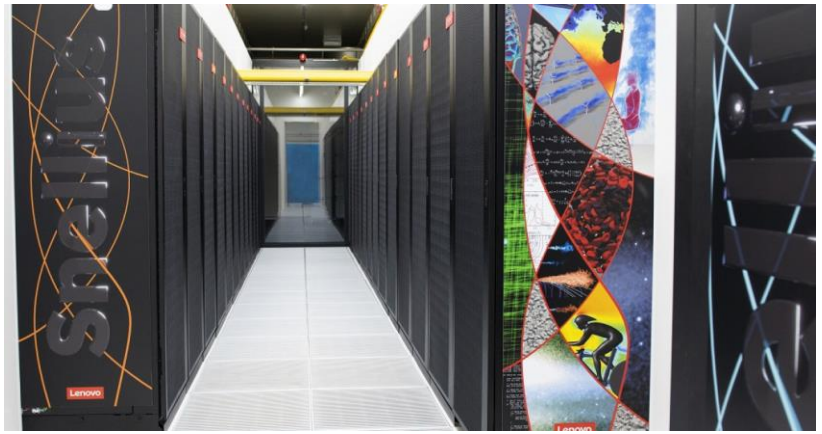
While…

- EU organisations have obtained licenses for use of LLMs Big Tech
- Publishers try to negotiate license fees for use of their content for LLM training purposes
- LLM **training capacity** of U.S. providers **may be claimed by U.S. Government**
- License payments are used for further AI investments (and not benefit EU innovation and economy)

Consortium

TNO innovation for life

SURF



Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

WHAT

A RESPONSIBLE LARGE LANGUAGE MODEL BUILT FROM SCRATCH

Data
900 billion text tokens + 245 billion code tokens

- Opt-in data
- Data that is legally accepted for the training of LLMs
- Non-IP infringing synthetic data

Performance
Comparable to the Llama2 7B model, GPT-3 175B models

- Text generation
- Summarization
- Simplification

Reciprocity

We believe innovation should benefit everyone and should contribute to a fair and inclusive society. Technology should be built in cooperation with important stakeholders.

In current LLMs, **privacy, data and IP** is not enough protected. GPT-NL will be **built from scratch** in accordance with **AI Act, GDPR & IP law**.

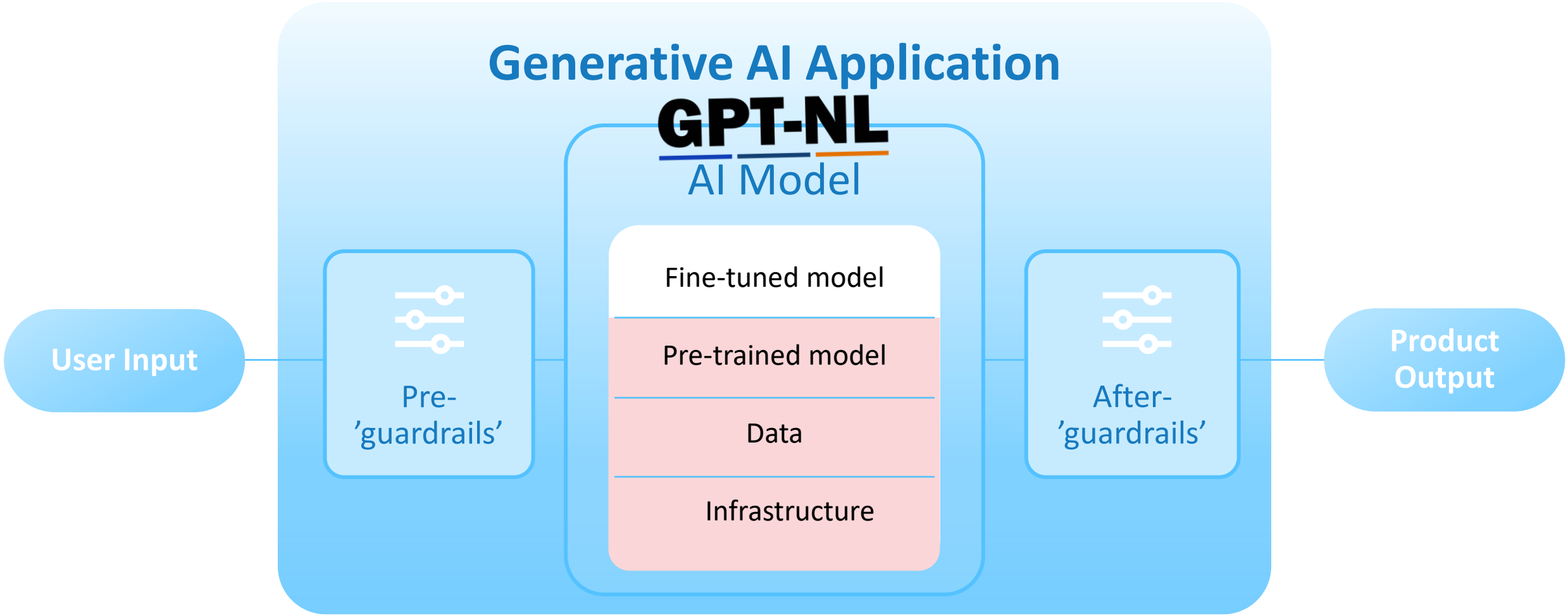
GPT-NL

Purpose-built for compliance: Designed to meet Dutch and European regulatory standards, ensuring legal robustness in AI applications.

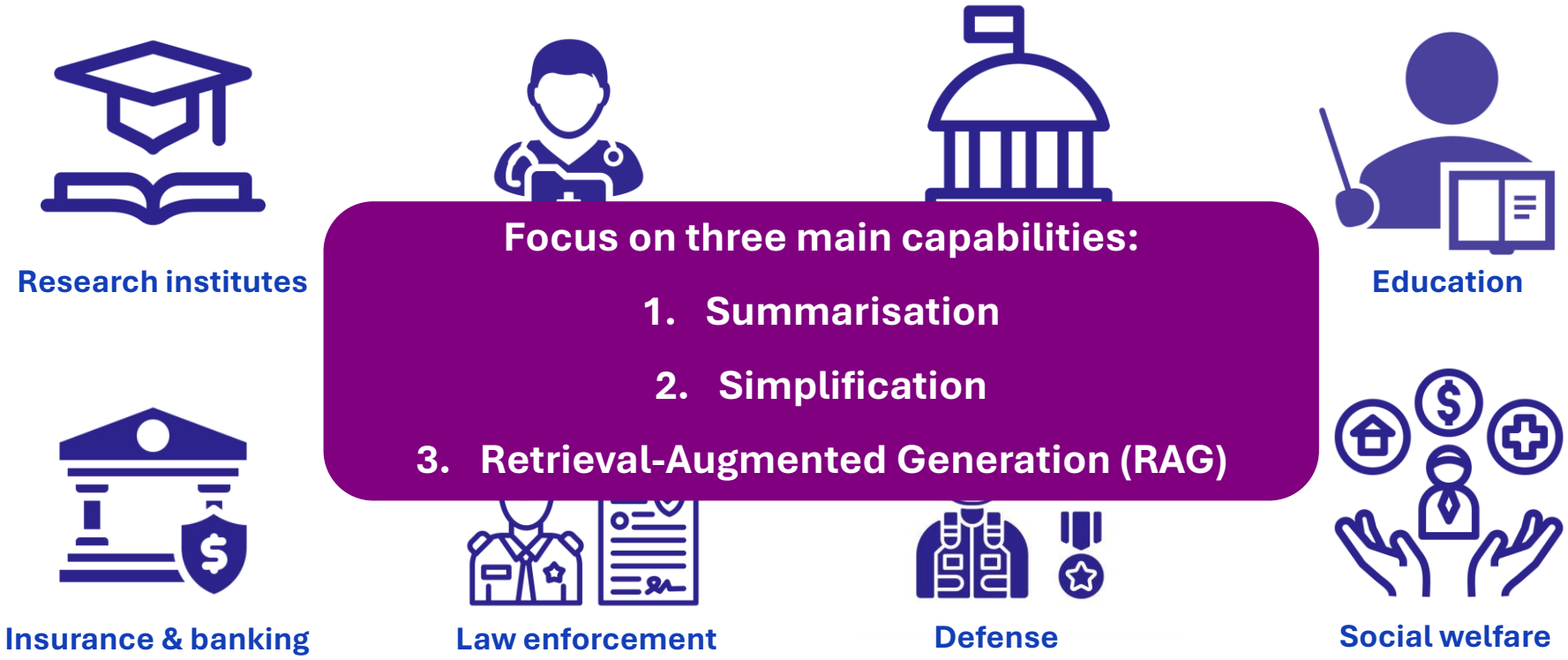
Trusted for high-stakes environments: Prioritizes legally compliant data, making it ideal for applications and industries where compliance is paramount.

A foundation for innovation: Delivers sufficient performance for general tasks, providing a reliable base for domain-specific fine-tuning and further research.

Models are part of applications



Capabilities of GPT-NL



GPT-NL licensing

- Licensing types and conditions still work in progress!
- Component-based licensing

- **Source code** will be made available under open license
- **Datasets** mixed as some is public, some is proprietary
- Gated access to **model weights** for research purposes – research license
- Paid access to **model weights** for all other purposes – enterprise license

Milestones 2024

- Set-up data strategy
- Kick-off Content Board external data contributors
- Data curation pipeline completed
- Model architecture and training framework: compared frameworks for efficiency gains

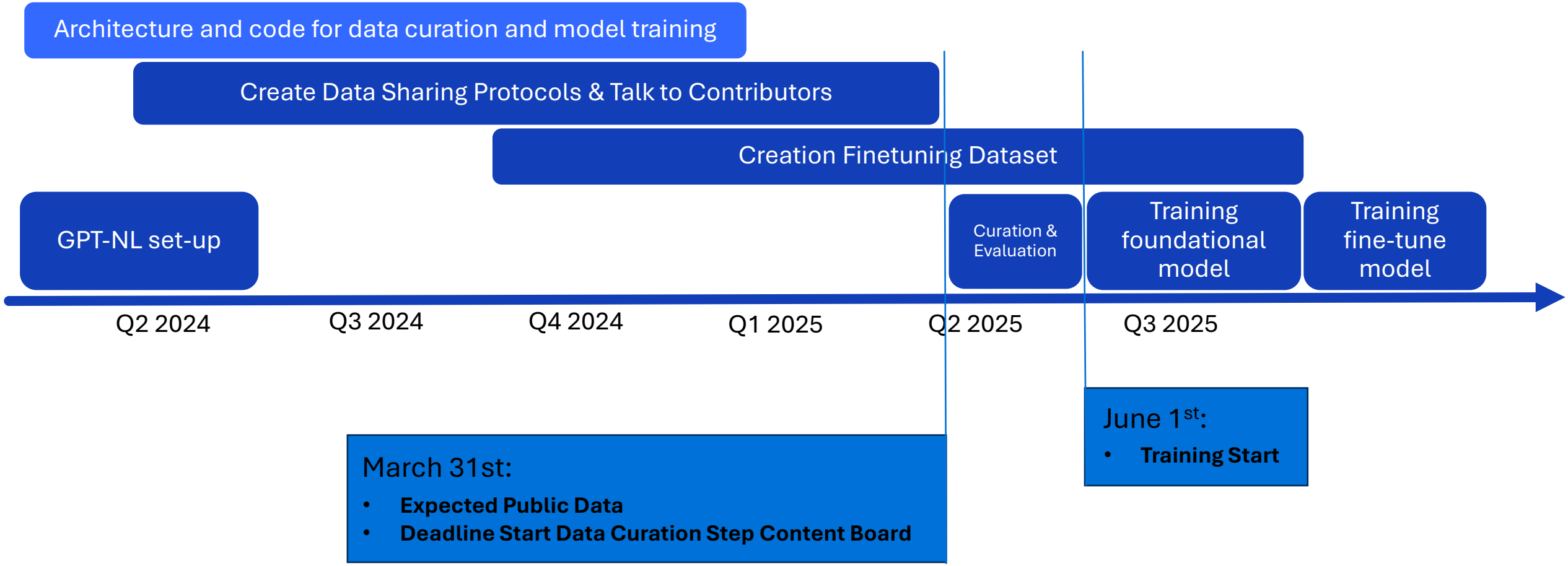
- And…

- Lots and lots of legal discussions

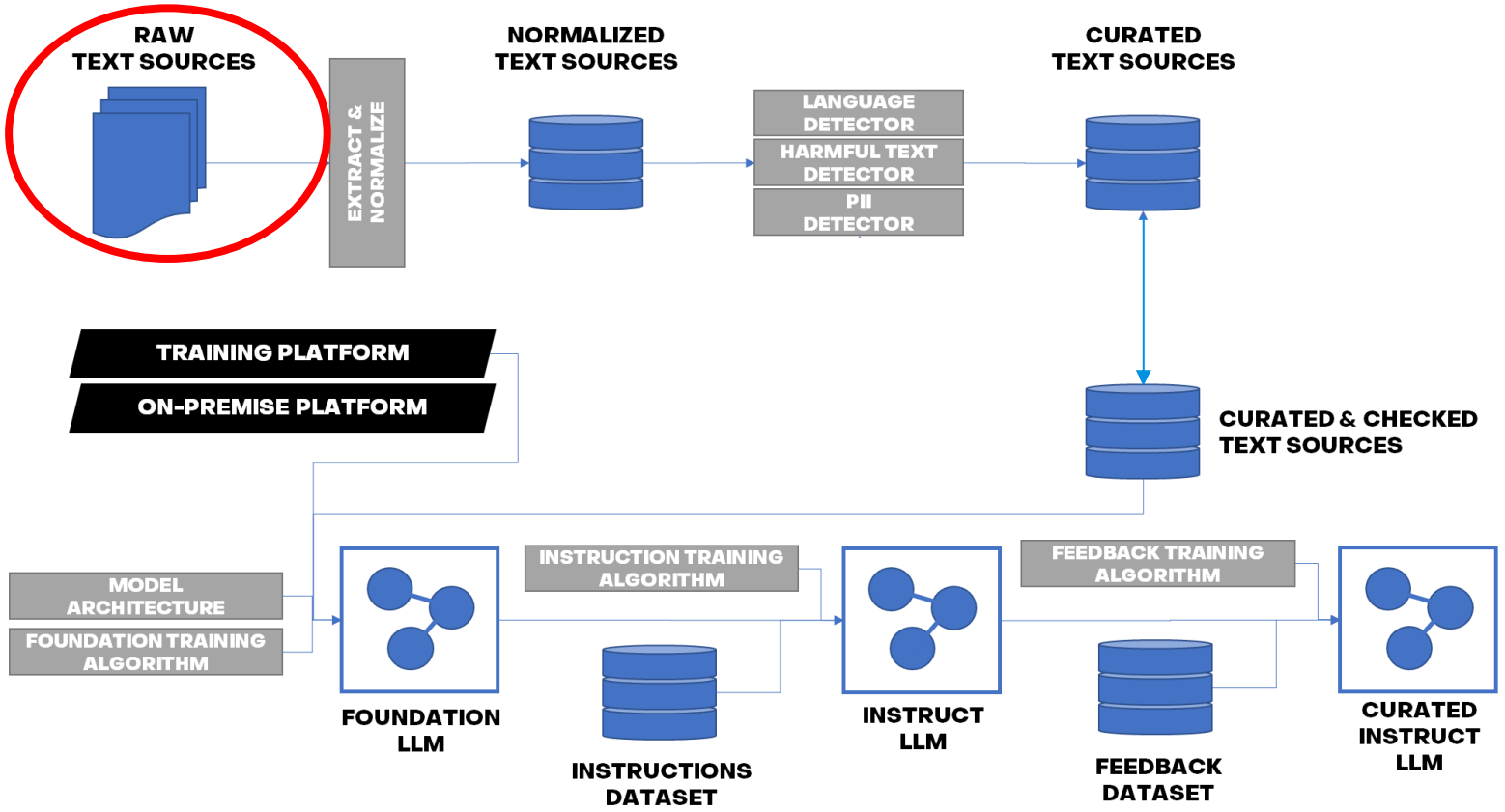
There is no reason AI can't be compliant with GDPR, but companies need to take the time to get it right... Organisations need to prioritise legality over speed. After all, the backlash over a legal issue is much more significant than that of the potential complaints over the timeline.

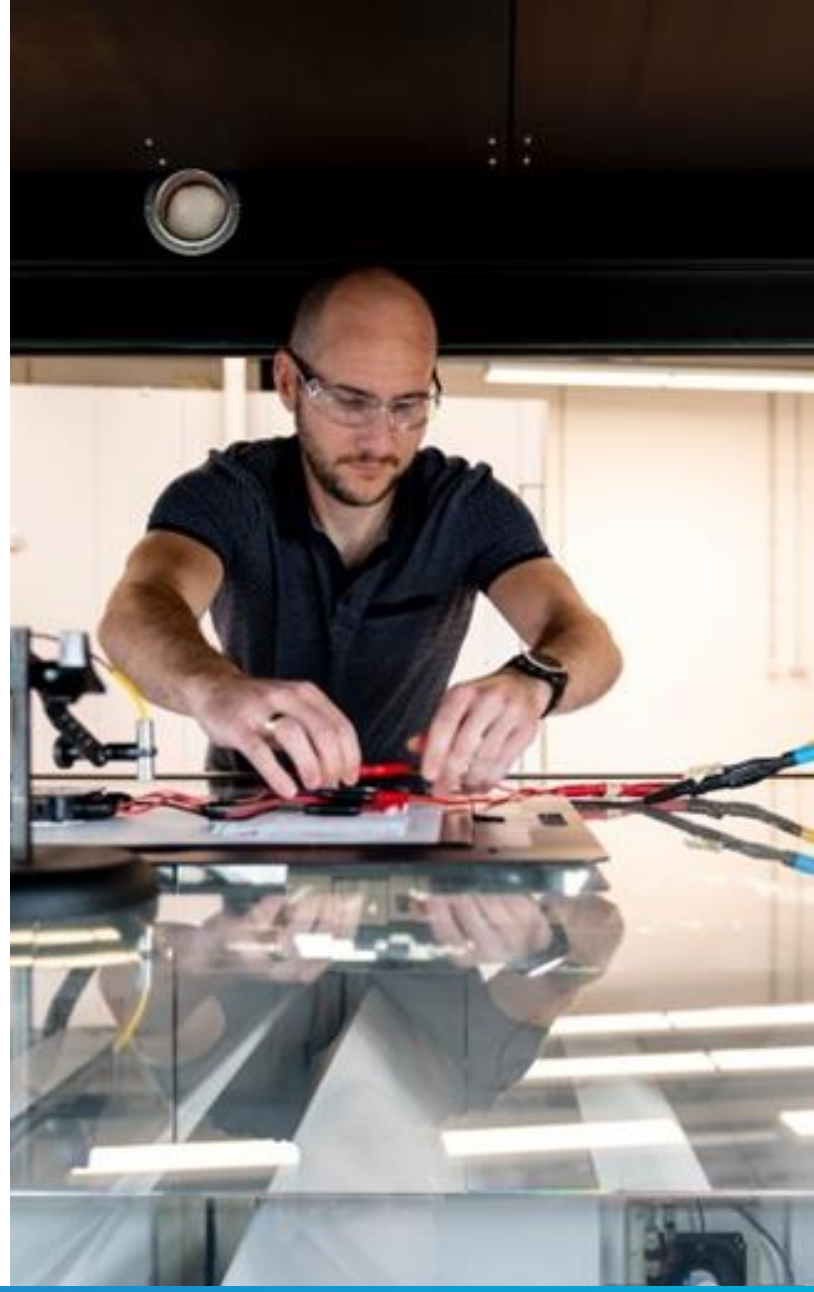
Chris Denbigh-White

Planning


Architecture and code for data curation and model training
Create Data Sharing Protocols & Talk to Contributors
Creation Finetuning Dataset
GPT-NL set-up
Curation & Evaluation
Training foundational model
Training fine-tune model
Q2 2024
Q3 2024
Q4 2024
Q1 2025
Q2 2025
Q3 2025
March 31st:
• Expected Public Data
• Deadline Start Data Curation Step Content Board
June 1st:
• Training Start

HOW?


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



Datasets for GPT-NL (in billion tokens)

~36B

Proprietary data

~800B

Publicly available data

~40B

Synthetic Dutch data

~245B

Code

Proprietary datasets

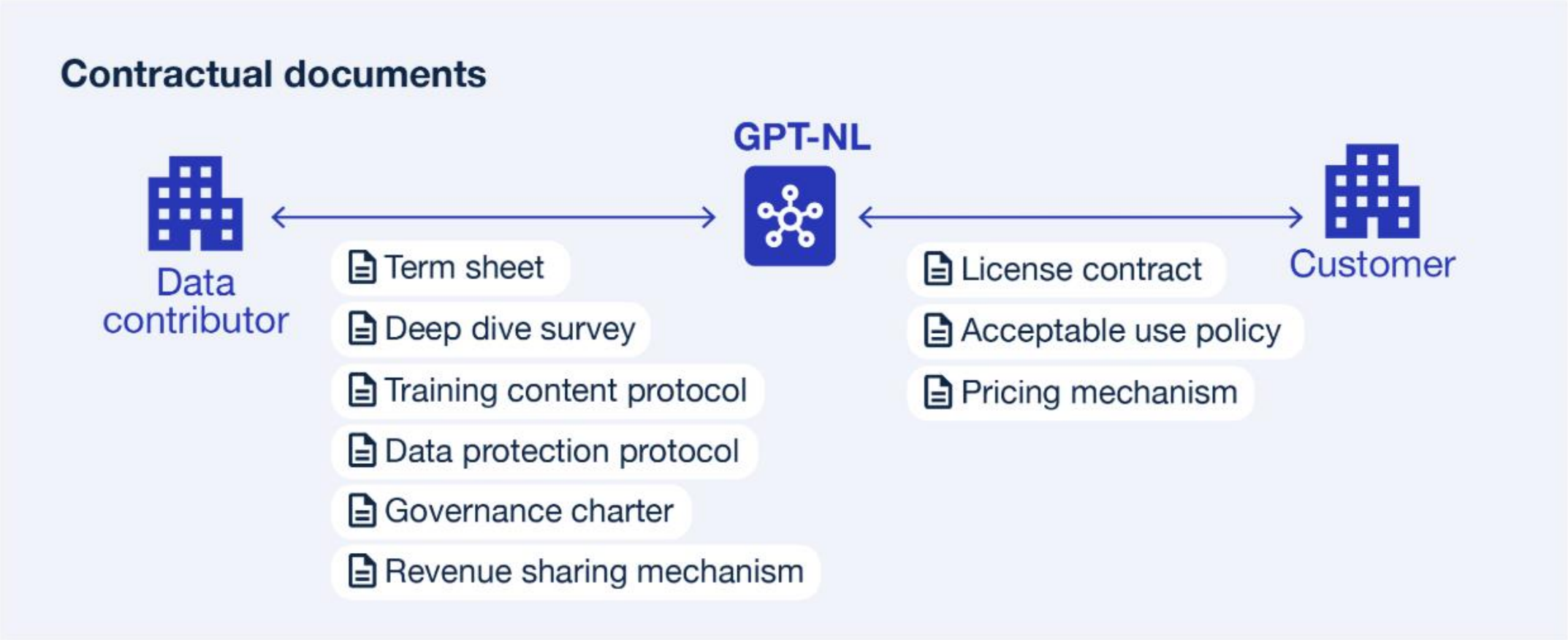
What's in it for external data contributors?

- Contribution to Dutch & fair Gen AI ecosystem
- Curated dataset – free of charge
- Commercial revenue share and/or discount on license to LLM
- Better performance of the LLM for their use-case

Building trust

- Organisations in the content contributor board trust us as we've been engaging in dialogue instead of trying to use their data unlawfully
- We now have data terms that at least a couple of big parties with varying backgrounds agree on.
- We have established something we can use to move forward in LLM development beyond GPT-1.0 with mutual trust.

Creating a uniform contract



- Website: https://redactie-tno-subsites.iprox.nl/gptnl/gpt-nl-visual-overview/?reload=true

Public data: Crawling, annotating, and re-using

With Public data we mean CC-0, CC-BY, or public domain datasets from parties we are not in direct contact with

Common Corpus V2

- largest public domain dataset released for training LLMs
- multilingual, including Dutch
- **selecting permissively licensed data**

Scraping with permission

- Collaboration with Open State Foundation
- **unlock data from public organizations**
- Crawling e.g. officielebekendmakingen.nl, openraadsinformatie.nl, public domain information from Koninklijke Bibliotheek, reports from PBL, papers from Naturalis, EP, ..

Subset Common Crawl Data

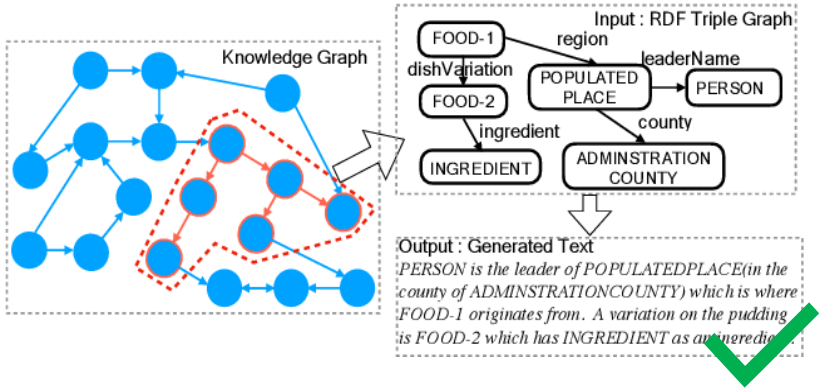
- collaboration with Bram Vanroy | Instituut voor de NL Taal
- **annotate Common Crawl data** with creative commons licenses
- identify licenses present in webpages
- listing domains that appear often in CommonCrawl, but where license is unclear

Synthesized data

TYPE I

Doing synthesis from content that we are licensed to use, even with using an LLM as a postprocessor, e.g. translating datasets.

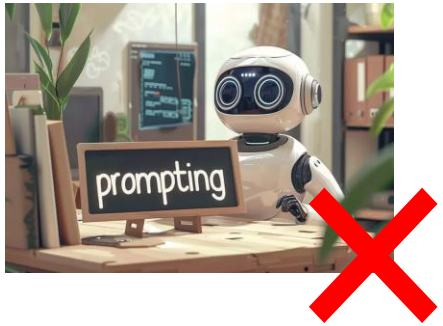
- **Low legal risk**



TYPE II

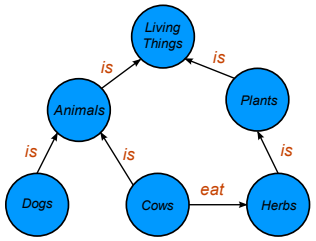
Using an LLM to directly generate content by prompting.

- **High legal risk**
- **In our view, not in line the GPT-NL ambition**



Data Synthesis type 1

Type 1 Data: Doing synthesis from content that we are licensed to use, even with using an LLM as a postprocessor

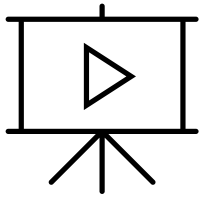


Graph

Low quality

Different language

CSV



Video

Done ✓

~3B tokens

Experimenting

Expecting nothing

In-Progress

Expecting

~40B low quality

Experimenting

Expecting nothing

Experimenting

Expecting nothing

Medium-Quality Dutch

GPT-NL

We can’t do it alone!

GPT-NL requires collaboration, honesty, and open discussion.

We’d love to hear from you if:

- You have any ideas for a strong, sovereign AI ecosystem within the Netherlands;
- Can help us in getting a rich, diverse dataset. Only together we can build GPT-NL!

Contact: info@gpt-nl.nl or follow us on LinkedIn!

edu.nl/mgvd3