GPT-NL

Considerations when creating the GPT-NL Corpus

Jesse van Oort

CLIN - Leuven

12-09-2025

Why GPT-NL

nieuwsuur

Donderdag 22 mei, 17:03

Nieuw onderzoek: AI verbruikt 11 tot 20 procent van wereldwijde stroom datacenters

AI was in 2024 naar schatting verantwoordelijk voor zo'n 11 tot 20 procent het wereldwijde stroomverbruik van datacenters. Dat blijkt uit nieuw Neder onderzoek. Uit eerdere prognoses van het Internationaal Energieagentsch (IAE) en Netbeheer Nederland blijkt dat ook het stroomverbruik van datace door AI in de toekomst enorm gaat stijgen.

Vrijdag 3 mei 2024, 07:00

Chatbots recommend disinformation and fea mongering, tech companies tighten restricti

Fleur Damen
redacteur-verslaggever
Nieuwsuur

Roel van Niekerk
redacteur-verslaggever
Nieuwsuur

Google and Microsoft are limiting the answers their AI chatbots provic response to queries about the European elections. Their move follows investigation by *Nieuwsuur*, which found that the chatbots provided a violating their own policies and promises.

Menu | nrc | Abonneren | Podcasts | Digitale krant

Anthropic Agrees to Pay $1.5 Billion to Settle Lawsuit With Book Authors

The settlement is the largest payout in the history of U.S. copyright cases and could lead more A.I. companies to pay rights holders for use of their works.

NOS Nieuws • Zaterdag 27 juli 2024, 13:00

Waarom nieuwe AI-technologie voorlopig aan de EU voorbijgaat

Heleen D'Haens
correspondent Europese Unie

toezichthouder staat dat deze technologieën de kansenongelijkheid in de klas

What are we building?

- Is trained on data across the internet; making it a model that “knows everything”.
- Aims to be strong at every domain

BUT

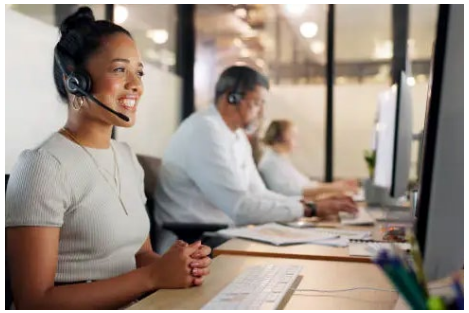
- Bad for cases with high requirements for data privacy, transparency, robustness, compliancy.
- Training models on the full internet is wasteful and introduces untrustworthy sources

- **Information should come from verifiable context**
- Training is focused on a limited set of tasks and domains that are most **useful** in our **Dutch critical sectors**.

- **Build with the AI Act in mind**
- Data put in the model is useful and of high quality, **using less untrustworthy sources**.

LLMs are good at interpretating and generating language, but not necessarily at answering questions without trustworthy context. GPT-NL is not trying to be the next google search replacement.

For whom are we building?



Services: Financial, Legal, Insurance, Telecom etc.

Government & Social welfare

Health Care

Research

Education

Safety, Security & Defence

Industry

To facilitate building and researching on the GPT-NL model, we need to be **trustworthy** and **transparent**

Trustworthiness in Data Collection

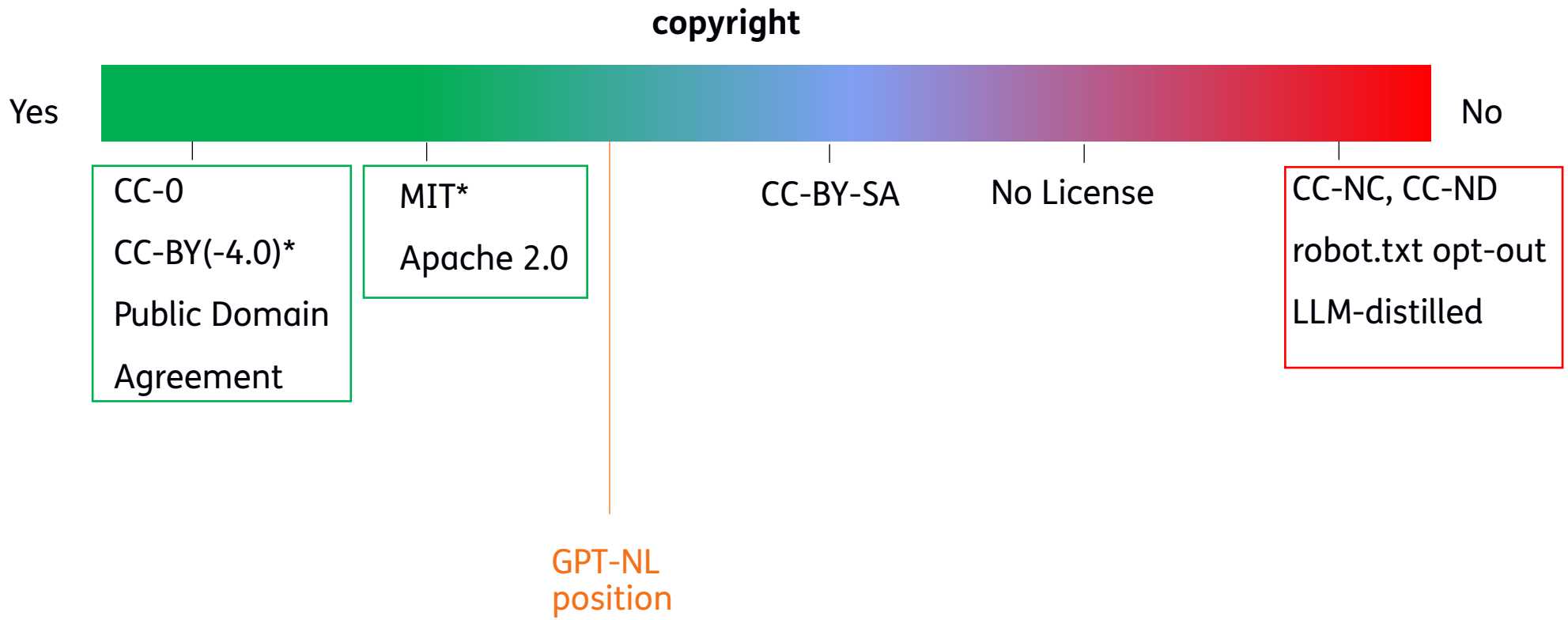
- Consent is required for data we train on
  - Consent through permissive open licensing; or
  - Through explicit (written) consent
- Collected data must help with achieving the main capabilities of the final GPT-NL model
- No confidential data and extensive removal of personal data
- Attribution for authors in CC-BY datasets

Trustworthy

Lawfulness (IP, Privacy), utility, committed, reliable

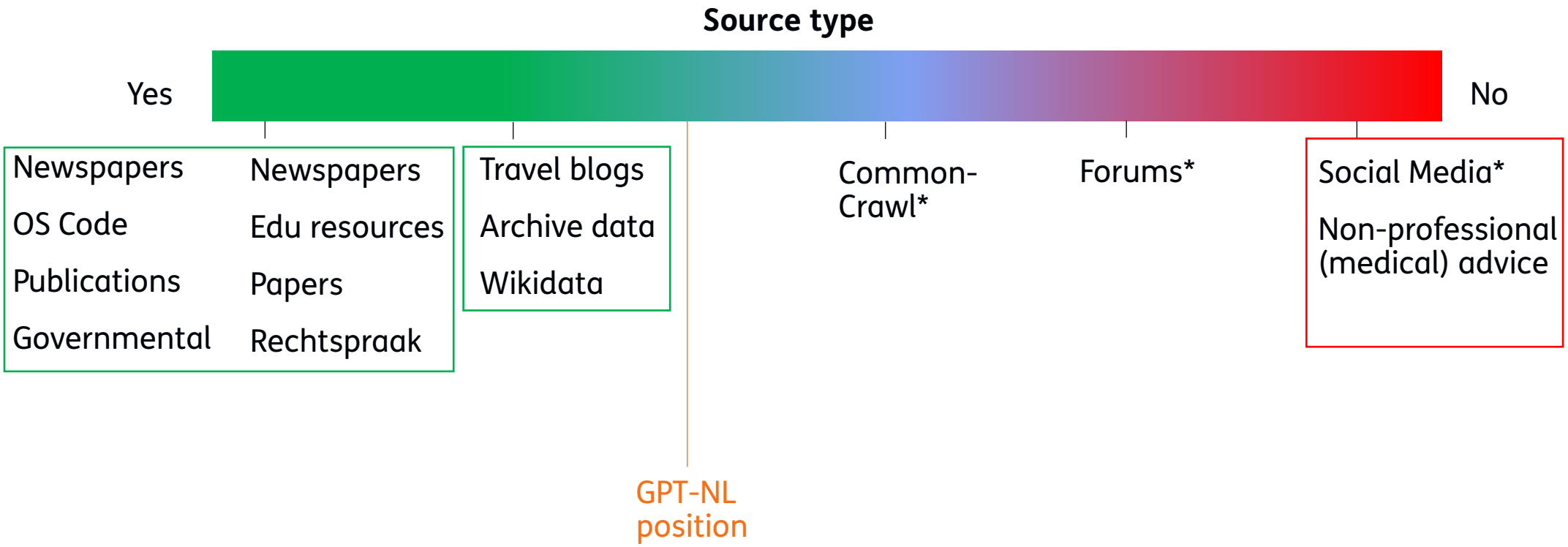
Data Guidelines

*without thourough curation



Data Guidelines

*without thourough curation



Selecting Existing Datasets

Creating/Collecting New Datasets

GPT-NL Corpus Acquisition Activities

Filtering Existing Datasets

Collaborating for Datasets

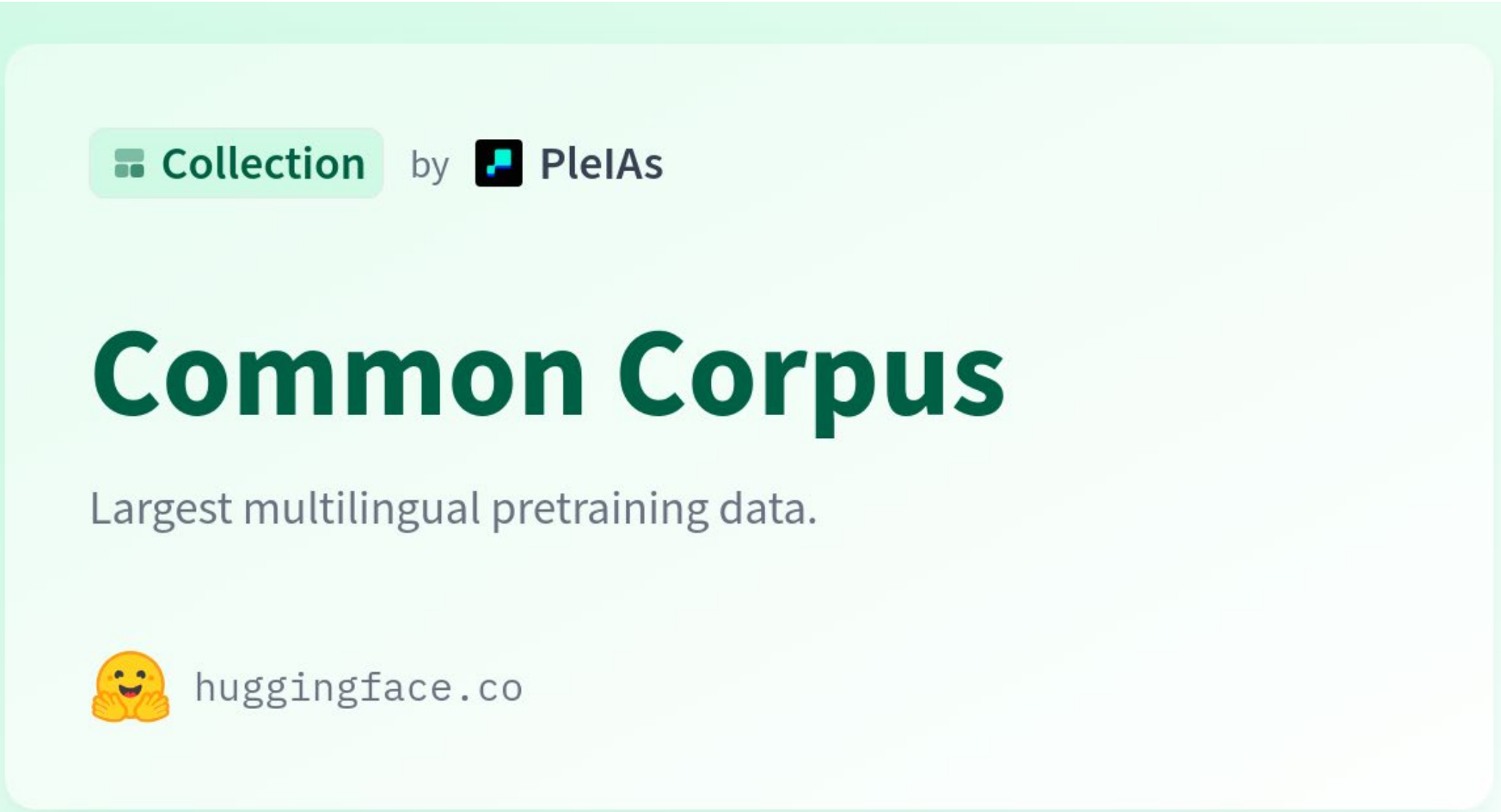
Data Selection

CC-0
CC-BY(-4.0)*
Public Domain
MIT
Apache 2.0

CC-BY-SA
CC-NC,
CC-ND
GPL-2.0,

Dutch
English
German
...
Greek
Spanish

Data Selection



CC-0
CC-BY(-4.0)*
Public Domain
MIT
Apache 2.0

CC-BY-SA
CC-NC,
CC-ND
GPL-2.0,

Dutch
English
German
...
Greek
Spanish

Data Selection

GPT-NL Corpus	Content	Q	NL (B)	EN (B)	Other (B)
Github Code	**Open Source Code**	**H**	**-**	**-**	**231**
OpenAlex	**Scientific research**	**H**	**0.06**	**47**	**0.4**
Eurlex	EU legal documents	H	0.09	0.08	0.33
English-PD	**English content in public domain**	**L**	**-**	**132**	**-**
German-PD	German content in public domain	L	-	0.3	31
Eurovoc	legislative documents from the EU	H	0.6	0.5	16
American-stories	**Historical American Newspapers**	**L**	**-**	**17.6**	**-**
Loc-PD-Books	**American open access catalog**	**L**	**-**	**7.5**	**-**

Other Sources →

Belgian Journal	Company bylaws	L	0.65	-	-
Rechtspraak	**Dutch court cases**	**H**	**2.3**	**-**	**-**

Creating and Collecting data

- Extracting existing datasources
- Creation of datasets by digitizing scans (e.g. Utrechts Archief)
- Creation of datasets by targeted scrapes with permission (e.g. Naturalis, Openraadsinformatie)
- Creation of synthetic data for permissively licensed WikiData graphs
- Translating Youtube Commons

These datasets and the tools/code necessary for creation will be made publicly available.

OPEN STATE FOUNDATION

YouTube

WIKIDATA

Creating datasets

GPT-NL Corpus	Work Done* (besides curation)	Content	Q	NL (B)	EN (B)	Other (B)
Utrechts Archief	OCR	Archive material	L	0.2	-	-
DPC/PBL/Tweedekamer/ Officiele bekendmakingen/Woogle	**API extraction**	**Government Communication Publications**	**H**	**6.7**	**-**	**-**
Openraadsinformatie	**API extraction**	**legislative documents from municipalities**	**H**	**14.1**	**-**	**-**
Naturalis	Scraping	Natural historic data	H	0.02	0.12	0.01
Youtube Commons	Translation	CCBY Youtube transcriptions	L	6.2	-	-
WikiData	Text Synthesis	Wikidata triples transformed to text	M	1.3	-	-

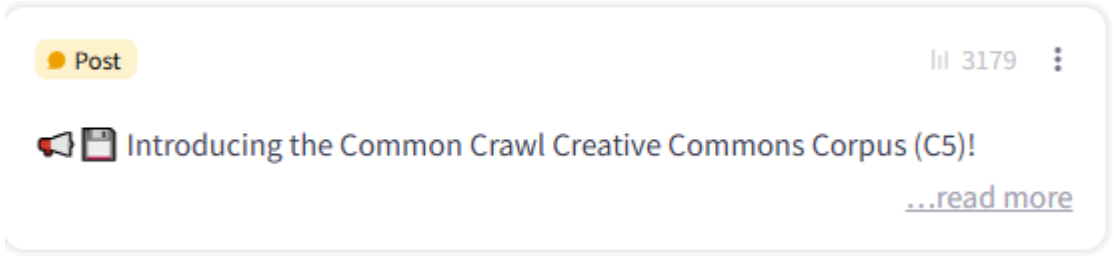
Collecting datasets

GPT-NL Corpus	Content	Q	NL	EN	Other
Nationaal Archief	Archive material	L	1	-	-
Auditdienstrijk	Audit reports by government	H	<0.01	-	-
DANS-KNAW	Research data	H	0.02	-	-
Koninklijke Bibliotheek	Dutch public domain	M	2.3	-	-
KPN	Telecom community	M	<0.01	-	-
Wikiwijs	Educational content	H	0.03	-	-
Zeeuwsarchief	Archive material	L	0.03	-	-
Noord-hollands archief	Archive material	L	0.02	-	-

Filtering existing data – C5

- Collaboration with Instituut van de Nederlandse Taal – KU Leuven (Bram Vanroy)
- Created a tool to annotate Common-Crawl with license data
- Manual false positive check

C5 dataset and tool is publicly available



Common Crawl Creative Commons (C5)

Language	No. Docs (original)	No. Docs (C5s)	No. Tokens (original)	No. Tokens (C5s)
afr	312,262	350	358,873,448	913,178
fry	230,910	1	197,430,774	1092
nld	2,827,636	18,266	2,757,074,705	18,957,134

Q	NL	Fry	Afr
M	0.04	-	-

Collaborating for data: the Content Board

GPT-NL

Home Contact

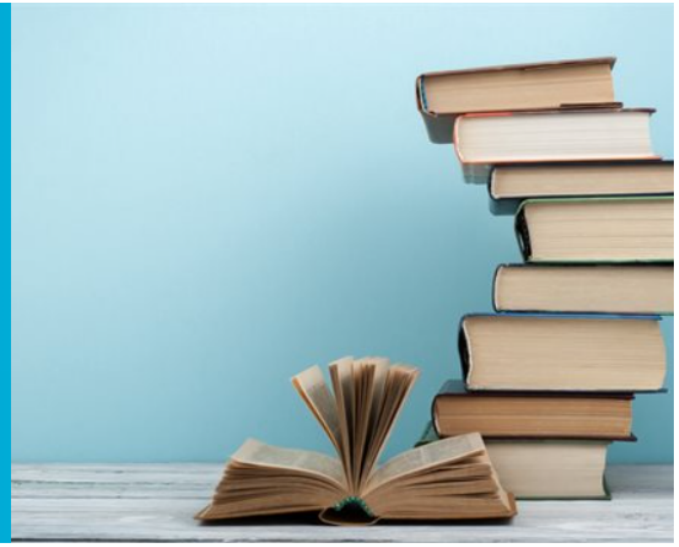
Over GPT-NL | Commitments | Samenwerken | Planning | Veelgestelde vragen | Nieuws

Home / Samenwerken / Content Board

Content Board

De Content Board bestaat uit de data providers voor GPT-NL, zie het als een soort vereniging of belangenvertegenwoordiging van de auteursrechthebbenden die hun data ter beschikking hebben gesteld voor het trainen van GPT-NL. Vanuit de Content Board hebben auteursrechthebbenden inspraak over de toekomst van GPT-NL. Zo, zetten we samen een stap in de richting naar een eerlijker AI innovatielandschap.

Op deze pagina vind je informatie over auteursrechtelijkbeschermde datasets, en informatie voor auteursrechthebbenden.



Samen bouwen

Voor het trainen van GPT-NL is een enorme hoeveelheid data nodig die divers genoeg is om tot een inclusief en sterk taalmodel te komen. Echter is deze data beperkt beschikbaar. Content providers leveren dan ook waardevolle data, waarmee we GPT-NL breed toepasbaar kunnen maken.

Content Contributer Agreement

Hierin staan de gedetailleerde afspraken die we met Data Providers maken, zoals afspraken over het aanleveren en opschonen van de content.

Download hier de Content Contribution Agreement (pdf, 511 kB)

Bijlagen

Download hieronder de bijgaande annexes van de Content Contributor Agreement.

Training Content Protocol (Annex 2) (pdf, 698 kB)

Governance Charter (Annex 3) (pdf, 263 kB)

Baseline Responsible Use Policy (Annex 4) (pdf, 169 kB)

Data Protection Protocol (Annex 5) (pdf, 255 kB)

Revenue Sharing Mechanism (Annex 6) (pdf, 1.6 MB)

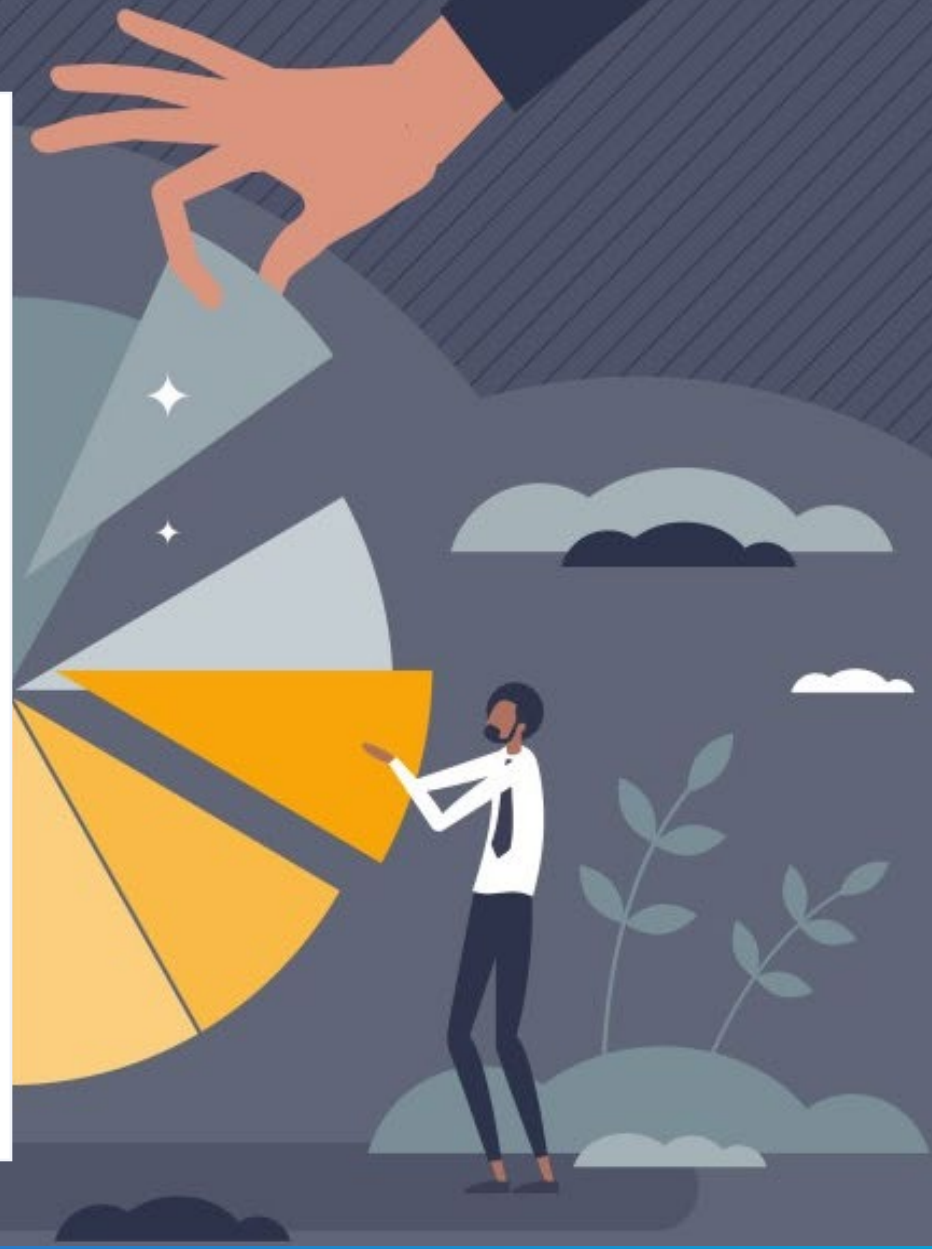
Reciprocity

Accessible for everyone, fair, diverse & inclusive

Read the full Content Contributor Agreement on www.gpt-nl.nl/samenwerken/content-board

Towards a fair data value chain

- **We believe data artists, journalists and other creators should be paid for their work**
- We **pay 50%** of the revenue from the commercial license to the owners of the data.
- The other 50% will solely be used for **continuation** of GPT-NL – **not for profit**.



Collaborating on data

GPT-NL Corpus	Content	Q	NL	EN	Other
NDP	**Umbrella dutch media**	**H**	**23.1**		
ANP	**Dutch news medium**	**H**	**0.97**	**-**	**-**
BNR	Dutch news medium	H	0.03	-	-
HBOs	Student thesis'	H	0.05	0.02	-
DNB	Financial documents	H	< 0.01	-	-
Centerdata	Societal research	H	< 0.01	-	-
ICTRecht	ICT law documents	H	< 0.01	-	-
Instituut van de Nederlandse Taal	Dutch linguistic research/data	M	0.92	-	-
Movisie	Social reports	H	< 0.01	-	-
Nederlandse tijdschrift voor de Geneeskunde	**Medical articles**	**M**	**0.49**	**-**	**-**
Waarbenjij.nu	Travel blogs	M	0.18	-	-
Fryske Akademy	Frysian linguistics	H	-	-	<0.1 (Fry)

GPT-NL Training Set

GPT-NL Corpus	NL (B)	EN (B)	Code (B)	Other (B)
High Quality	45	49	231	17
Low-Medium Quality	8.6	159	-	31

Finetuning efforts

#	Task	Created
3000	Closed question (based on context)	SpectrumAi
1500	Open question	
900	Brainstorm task	
1200	Chat conversations (5-7 messages deep)	
3000	Generating content (e.g. professional e-mail writing)	
900	Classification	
2250	Simplification (based on context)	
2250	Summarisation (based on context)	
18000	Open Question	GoeieVraag Onderdeel van Startpagina

GPT-NL Transparency Principles

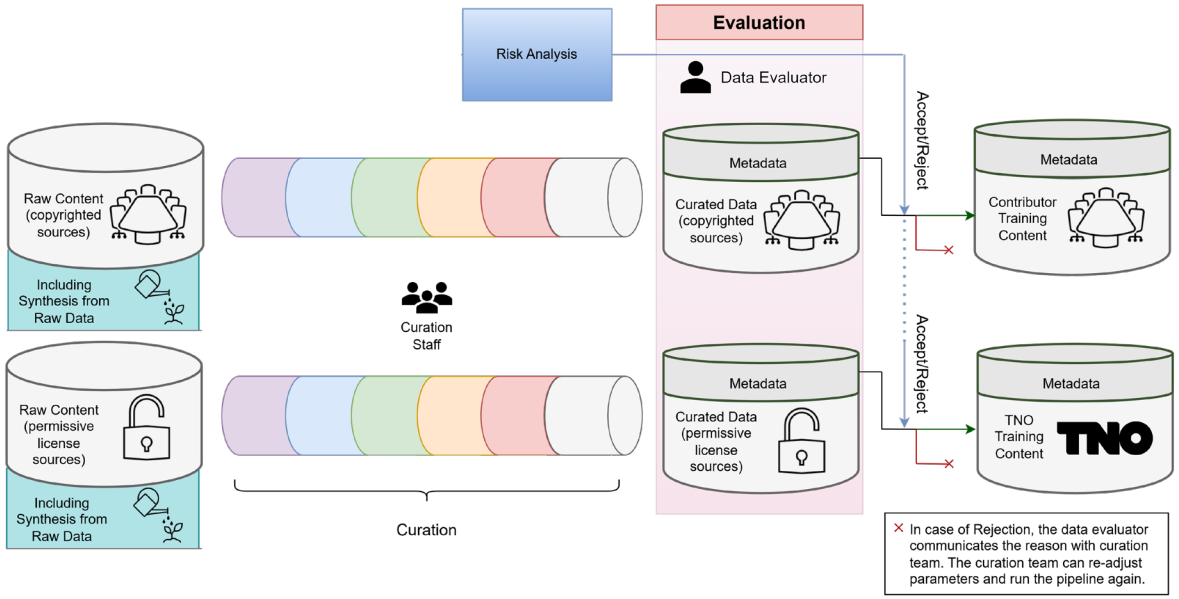
- Using uniform agreements for all stakeholders that can be seen by everyone
- Open-sourcing all data that we are allowed to publish
- Open-sourcing the curation pipeline and training code
- Release extensive metadata for all collections in the dataset
- Release statistics about training and inference, such as energy consumption
- Documenting and publishing key decisions

Transparency

Accountable, verifiable, honest

Transparency in the (non-public) Dataset in two ways

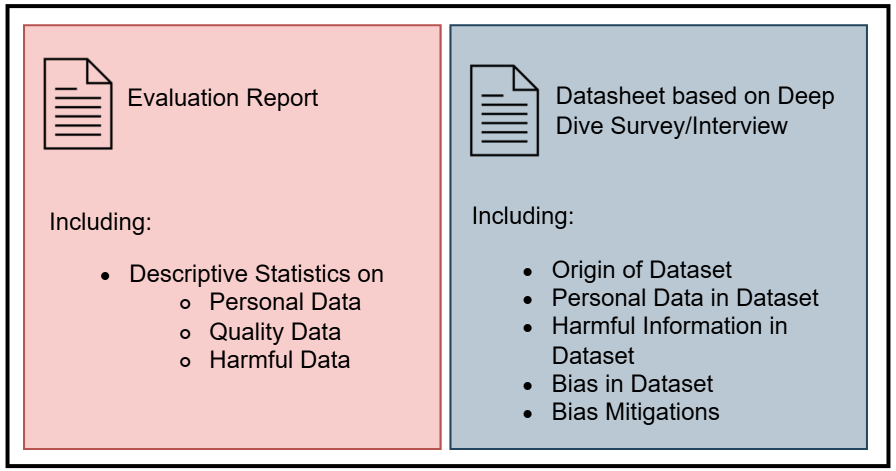
- Interviewing/surveying Data Contributors for appropriate meta-data
- Evaluation of received datasets



Metadata Field	Data Contributor Response
1.1 Description Organization Dataset	
2.1 Origin Content	
2.2 Topics	
2.3 Languages	
2.4 Quality-assurance	
2.5 Date Created	
2.6 Collection Methods	
2.7 Date Collected	
2.8 Author Dataset	
2.9 Available Metadata	
2.10 Modifications made	
3.1 Personal Data in Set	
3.2 Reason for Personal Data	
3.3 Accuracy Assurance Personal Data	
3.4 Request for Personal Data Removal Implementation	
3.5 Assurance of no confidential/protected information	
3.6 Harmful Data in Set	
4.1 Ethical Review	
4.2 Author Background	
4.3 Diversity Metadata	
4.4 Potential Biases	
4.5 Bias Mitigations	
5.1 Additional Information	

Transparency in the (non-public) Dataset in two ways

- Interviewing/surveying Data Contributors for appropriate meta-data
- Evaluation of received datasets
- Metadata for all sets will be published

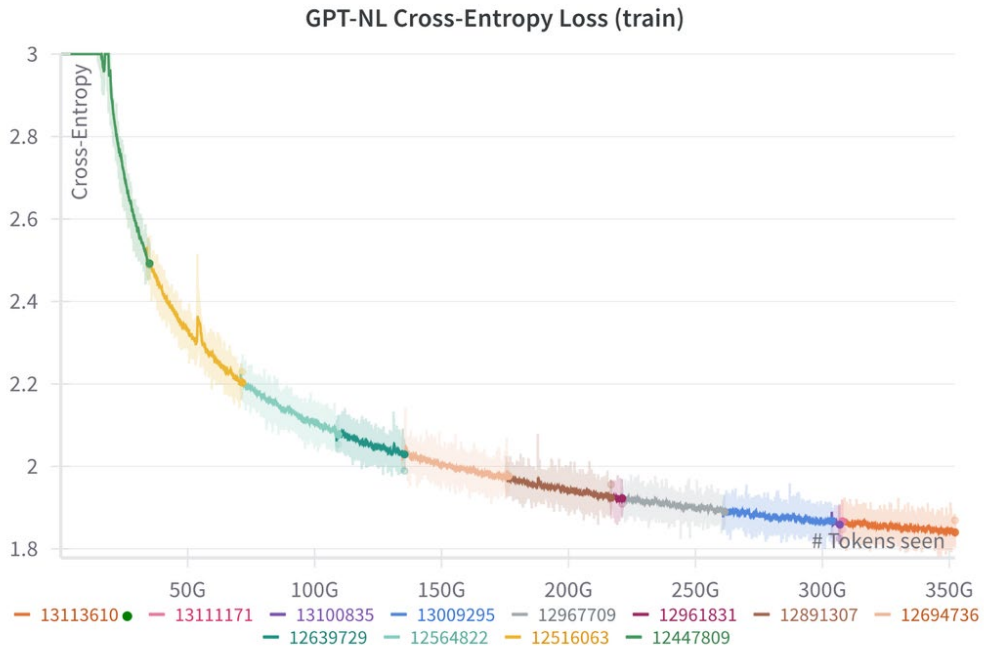


To-dos

- Benchmarking & Evaluation
- Publishing (meta)data
- Alignment tuning

And much more..

Snellius is working hard



After prompting: "Een aardappel is..."		
After 3B tokens	After 210B tokens	After 350B tokens
"...een belangrijk onderdeel van de voedselketen. Het gaat om het behoud van biodiversiteit en het vergroten van de weerbaarheid tegen klimaatverandering. De eiwittransitie moet leiden tot minder gebruik van gewasbeschermingsmiddelen, meer plantaardige alternatieven voor dierlijke mest..."	"...een waardige vervanger van de biet, en het is een goede keus om de teelt van aardappelen te combineren met andere teelten. De aardappel kan als basisgewas geteeld worden op alle grondsoorten, maar heeft een voorkeur voor..."	"...een plant, die in de grond wortelt en waarvan het loof dient als voedsel voor dieren. De aardappel is dus een gewas. Het telen van gewassen is een landbouw-activiteit. In de Noordoostpolder zijn twee..."

Stay in Touch!

Jesse van Oort

Data Acquistion Lead @GPT-NL

Responsible AI Researcher @TNO

jesse.vanoort@tno.nl