

GPT-NL

Recipe for training GPT-NL

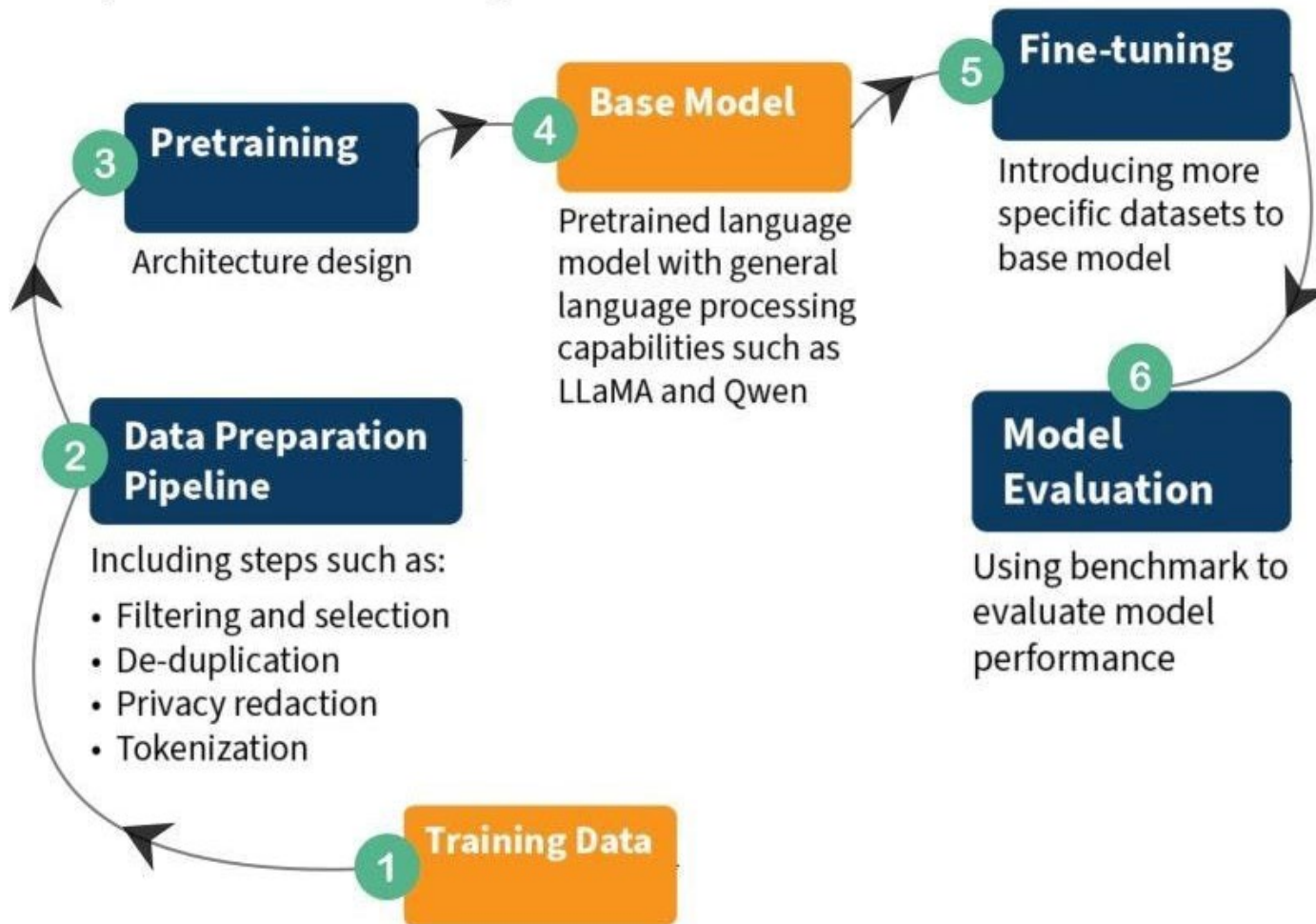
De Nederlandsche Bank



Jesse van Oort

- Researcher Responsible AI @ TNO
- Data Acquisition Lead @ GPT-NL

LLM Training Workflow (simplified)



**What are we
building?**

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 van het wereldwijde stroomverbruik van datacenters. Dat blijkt uit nieuw Nederlands onderzoek. Uit eerdere prognoses van het Internationaal Energieagentschap (IAE) en Netbeheer Nederland blijkt dat ook het stroomverbruik van datacenters door AI in de toekomst enorm gaat stijgen.

Vrijdag 3 mei 2024, 07:00

Chatbots recommend disinformation and fear mongering, tech companies tighten restrictions



Fleur Damen
redacteur-verslaggever
Nieuwsuur



Roel van Niekerk
redacteur-verslaggever
Nieuwsuur

Google and Microsoft are limiting the answers their AI chatbots provide in response to queries about the European elections. Their move follows an investigation by *Nieuwsuur*, which found that the chatbots provided answers 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

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 interpreting 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

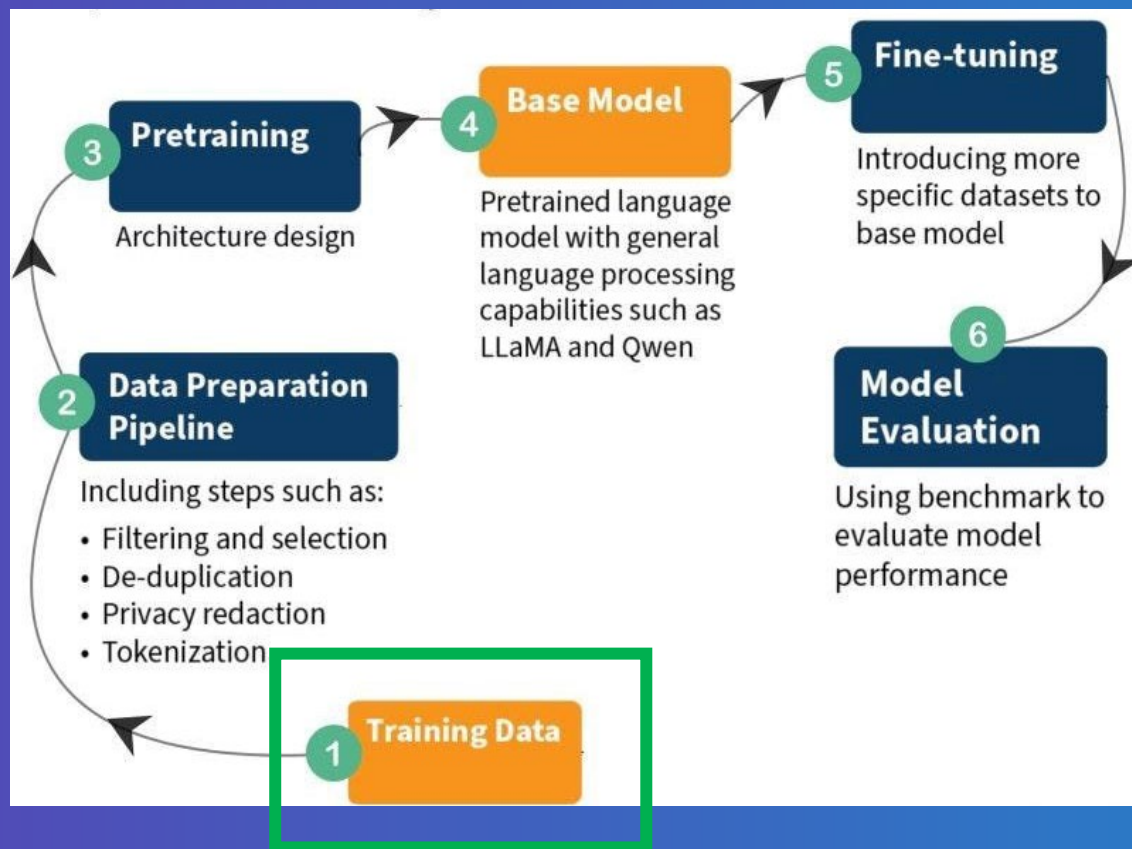


Safety, Security & Defence



Industry

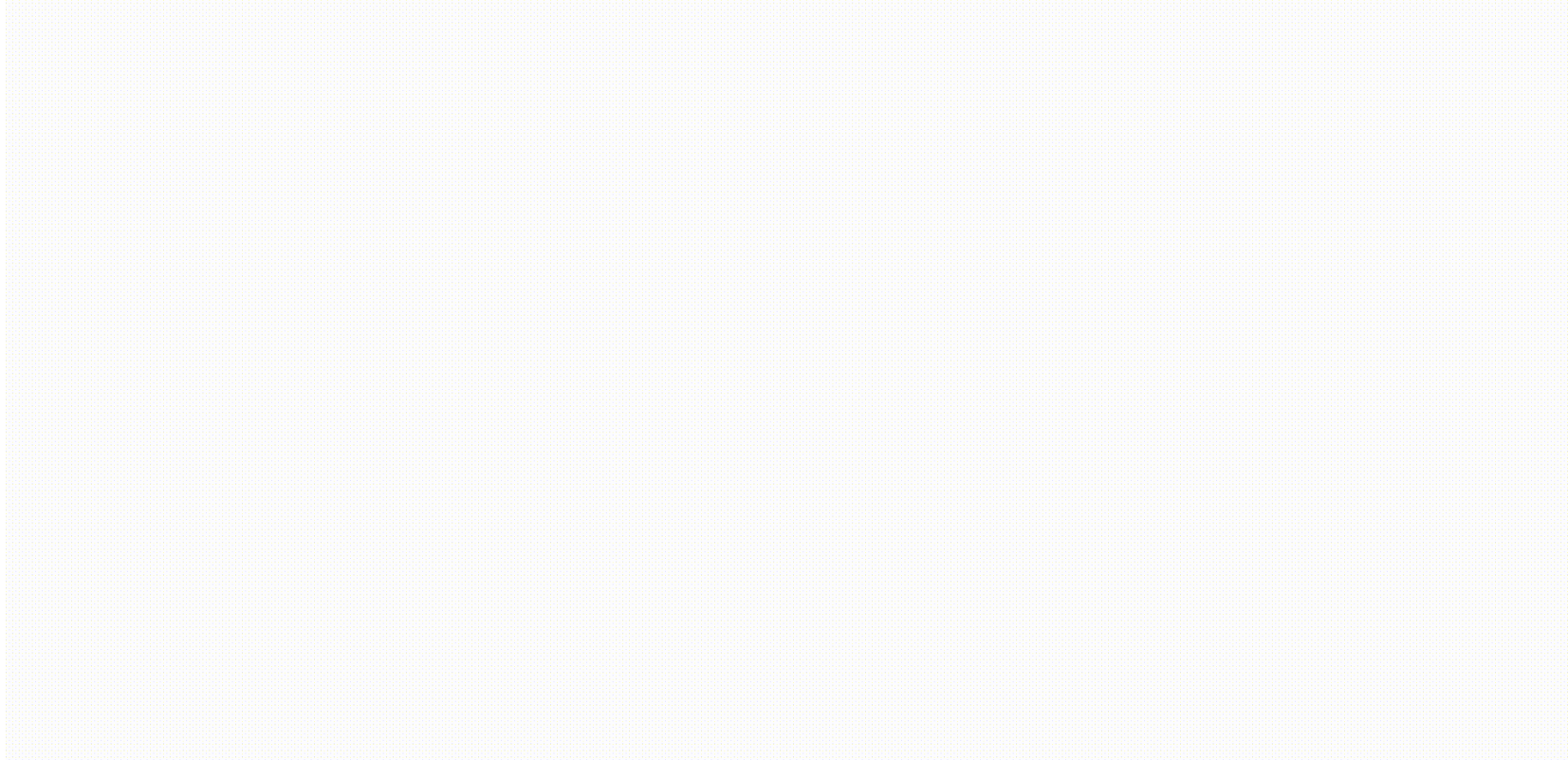
To facilitate building and researching on the GPT-NL model, we need to be **trustworthy** and **transparent**



Training Data

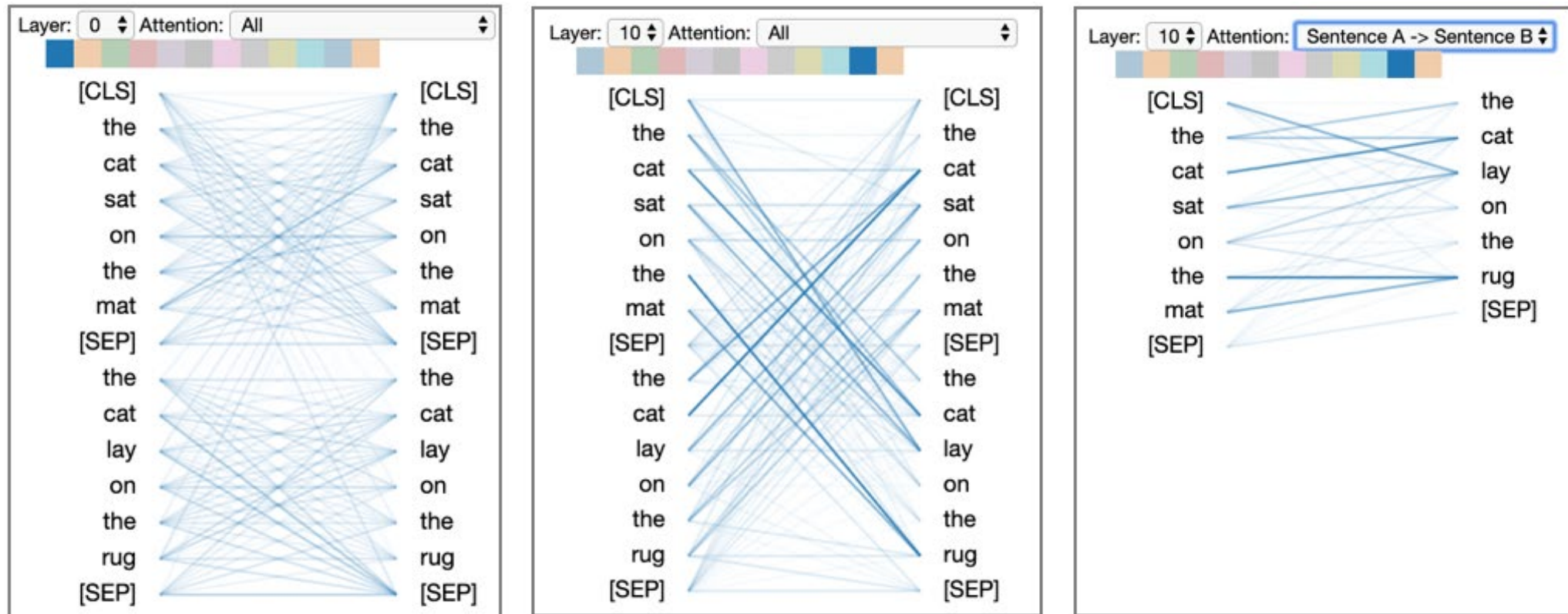
LLMs

- Recall: pretraining an LLM is simply next-token prediction



Data requirements

- Transformers, in particular LLMs, are data hungry!
- But why? Attention!



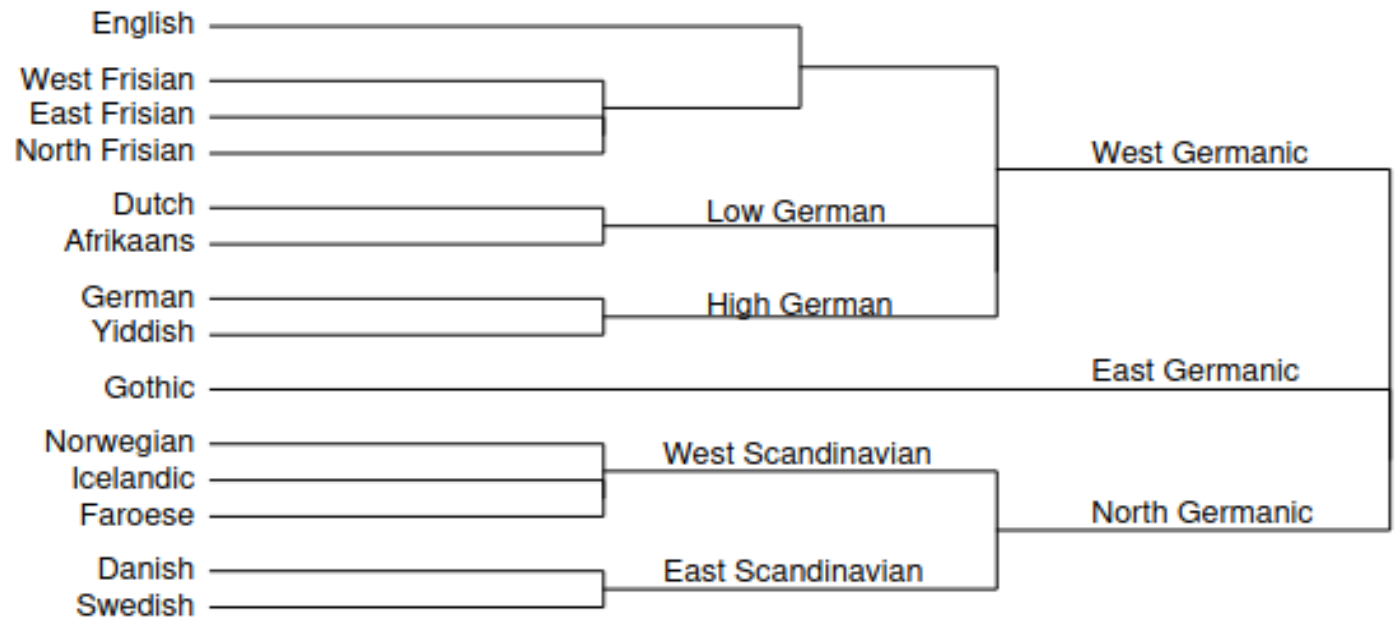
Source: J. Vig, *Visualizing Attention in Transformer-Based Language Representation Models*, arXiv:1904.02679, 2019.

Data requirements

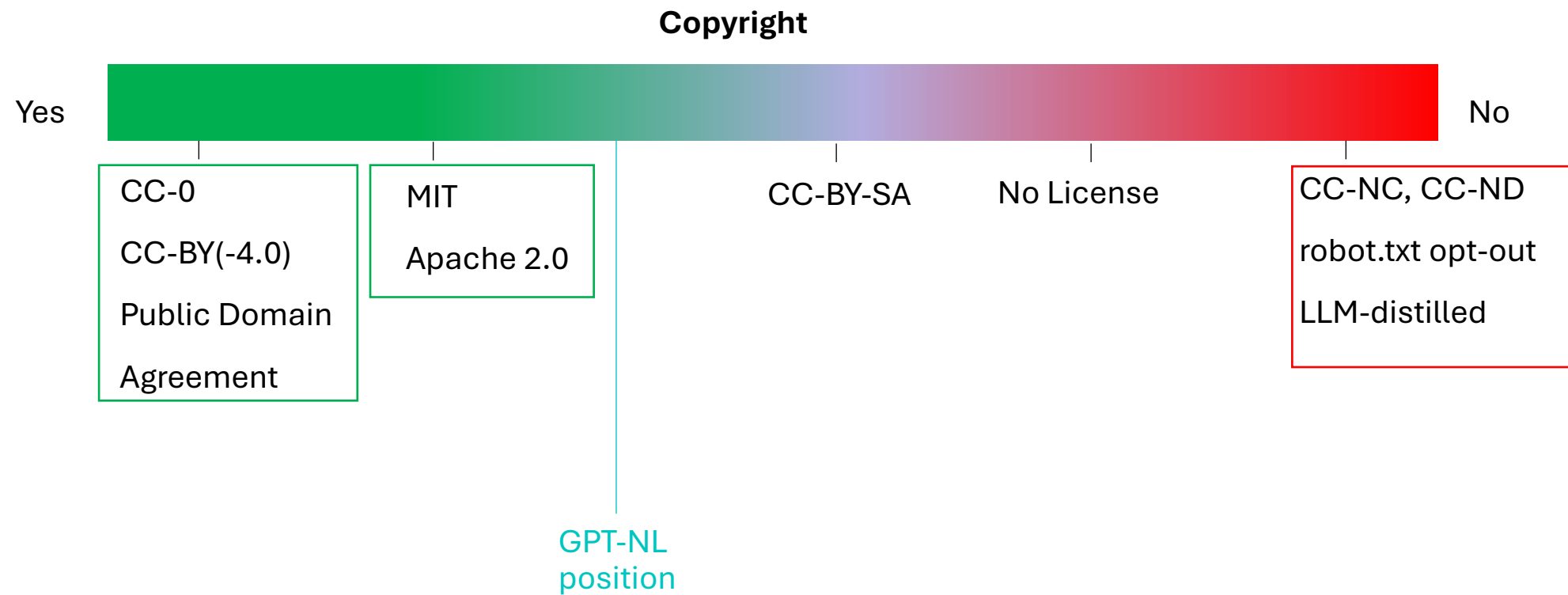
- LLMs are data hungry!
- Target is minimum **300B** *authentic* tokens (± 225 billion words), 150B (50B minimum) **Dutch**, 150B English
- There is simply not enough (Dutch) open data available to train a model. We need to do more.

Also English, because

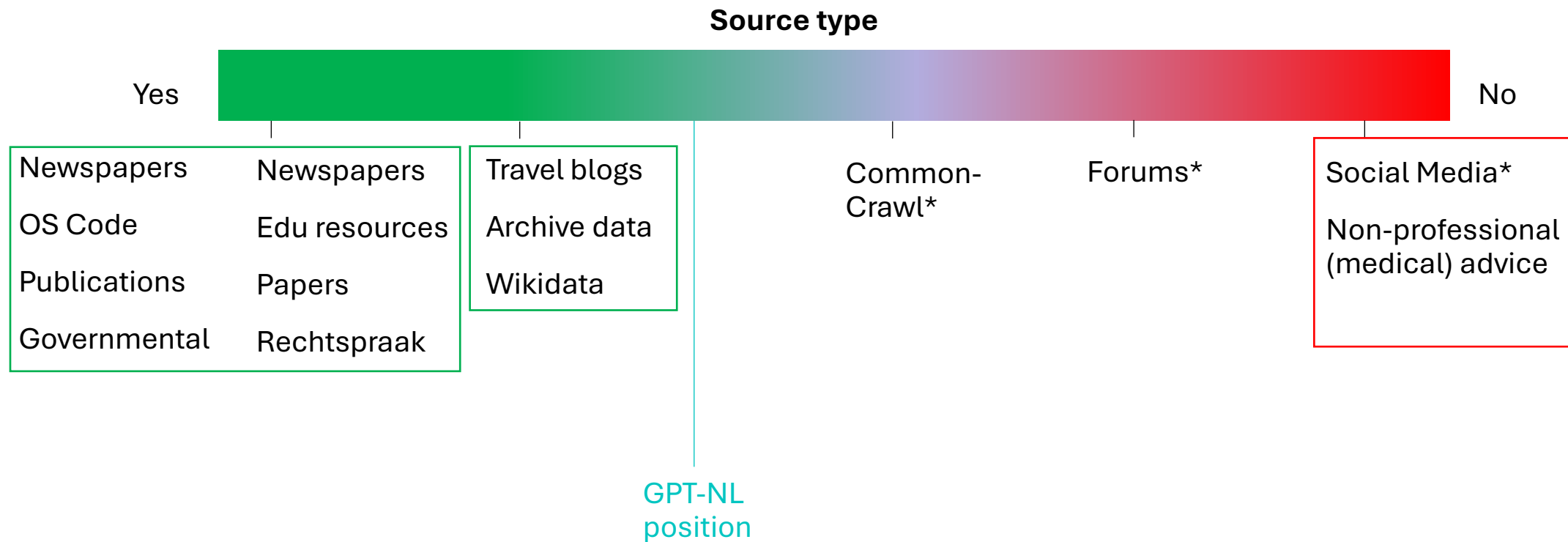
- English very similar to Dutch
- More tokens = **higher quality** responses
- Open data available in English
- Users will likely also be able to speak English to the model



Data requirements



Data requirements



**Without thorough curation*



Collecting Datasets

GPT-NL Corpus
Acquisition Activities

Creating New
Datasets



Selecting/Filtering
existing Datasets

Collaborating for
Datasets





Data Selection

 Collection by  PleIAs

Common Corpus

Largest multilingual pretraining data.



huggingface.co

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

 Collection by  PleIAs

Common Corpus

Largest multilingual pretraining data.



huggingface.co

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



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



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

[C5](#) dataset and tool is publicly available



Introducing the Common Crawl Creative Commons Corpus (C5)!

[...read more](#)



Data Selection/Filtering

Collection by PleiAs

Common Corpus

Largest multilingual pretraining data.

 huggingface.co



GPT-NL Corpus	Content / Source Description	NL (B)	EN (B)	Oth (B)
---------------	------------------------------	--------	--------	---------

Selected Data (GPT-NL Curated)				
Eurovoc	Multilingual EU vocabulary	0.6	1.3	16
Eurlex	EU law texts	0.09	0.08	0.3
OpenAlex	Academic corpus	0.06	48	0.4
English-PD	English public domain texts	0.02	132	0.5
American-stories	U.S. public domain literature	-	17.6	-
Loc-PD-Books	Library of Congress public domain books	-	7.5	-
Github Code	Open code data	-	-	232
German-PD	German public domain texts	-	0.3	31

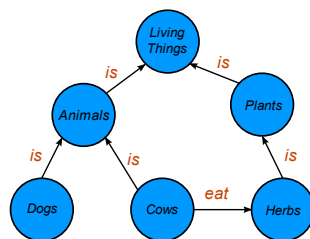


Filtered web-crawl Data (GPT-NL Curated)				
C5 Filtered	Web content	0.04	-	-



Creating Data

Source: Properly licensed but otherwise unusable data



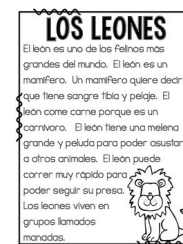
Graph



Tabular

tot komen, alwaar den Cardinael Porto
Carretero mede spoedigh nae toe is ver-
ken. Op de Begraefinge van de Marquis de Liches,
hebben haer, neffens meer ghevolgh, over de 200 Edel-
lyuden laten vinden. Den Prince Ludovico heeft de
Charge van Capiteyn onder het Regiment van de Quar-
de des Koninghe bekomen, vacant geworden zijnde, door
dien de Soos van den Cardinael van Monaco, die de
zeve plaats bekleeide, verkooren is tot Generael van de
Siciliaensche Galeys. Men segt nu dat den Marguis
de los Balbaes, tegenwoordige Gouverneur van Milan-
nen, tot Ambassadeur na het Hof van den Keyser is ge-
mineert, in de plaats van den Graef van Castigliar, ende
geordonneert soude werden sijne reysse derwaerts te nemen
jonder hier ten Hoof te komen.

Low quality



Different
language

High-
Quality
Dutch





Creating Data

- 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.

MADLAD-400

microsoft/phi-4



Para alguien que solo tiene un martillo en su caja de herramientas, cada problema se ve como un clavo.	Voor iemand die alleen een hamer in z'n gereedschapskist heeft, ziet elk probleem eruit als een spijker.
--	--



Triple	Willem-Alexander – noble title – Prince of Orange
Dutch Sentence	<i>Willem-Alexander heeft de titel Prins van Oranje.</i>

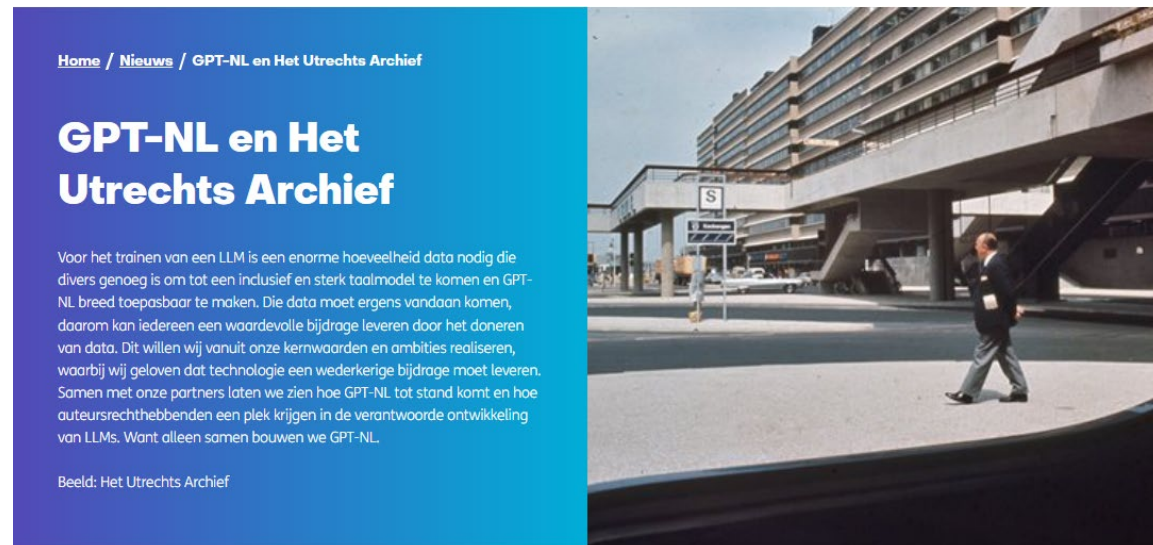


Creating datasets

GPT-NL Corpus	Content / Source Description	NL (B)	EN (B)	Oth (B)
Synthetic Data (GPT-NL Curated)				
Youtube-Commons-Synth	Public domain YouTube transcripts translated	6.2	-	-
WikiData-Synth	WikiData triples converted to running text	1.3	-	-

✓ Collecting Data

- Collaborating with public organisations
- Digitizing scans for archives
- Extracting distributed datasets with custom scraping tools.





Collecting datasets

GPT-NL Corpus	Content / Source Description	NL (B)	EN (B)	Oth (B)
Collected Data (GPT-NL Curated)				
Openraadsinformatie	Municipal council documentation	14.1	0.02	0.01
Officiële bekendmakingen	Government announcements	2.8	0.01	-
Woogle	Open Dutch government documents	2.6	0.14	-
Koninklijke Bibliotheek	Public domain Dutch texts	2.4	-	-
De Rechtspraak	Judicial cases	2.3	-	-
Tweede Kamer	Dutch parliamentary documents	1.3	-	-
Nationaal Archief	Dutch archive	1.1	-	-
Belgian Journal	Belgian company bylaws (Flemish focus)	0.7	-	-
Utrechts Archief	Dutch archive	0.2	-	-
Noord-Hollands Archief	Dutch archive	0.2	-	-
Zeeuws Archief	Dutch archive	0.2	-	-
Dienst Publiek en Communicatie	Dutch public communication docs	0.07	-	-
Wikiwijs	Dutch school content	0.03	-	-
PBL	Planbureau Leefomgeving docs	0.02	-	-
Naturalis	Biological publications	0.02	0.12	0.01
European Parliament	Multilingual EU documents	0.05	0.03	0.02
DANS-KNAW	Dutch archaeology descriptions	0.02	-	-
Auditdienst Rijk	Dutch audit publications	0.005	-	-

Collaborating for data: the Content Board

GPT-NL

[Home](#) [Contact](#) 

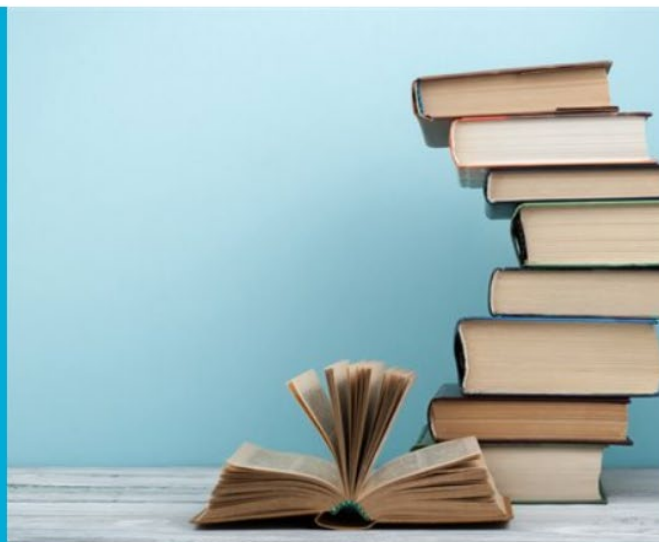
[Over GPT-NL](#) [Commitments](#) [Samenwerken](#) [Planning](#) [Veelgestelde vragen](#) [Nieuws](#) 

[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 auteursrechtelijk beschermde datasets, en informatie voor auteursrechthebbenden.



Reciprocity

Accessible for everyone, fair, diverse & inclusive

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 Contributor 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 \(Annex 6\) \(pdf, 1.6 MB\)](#) 

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

GPT-NL

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

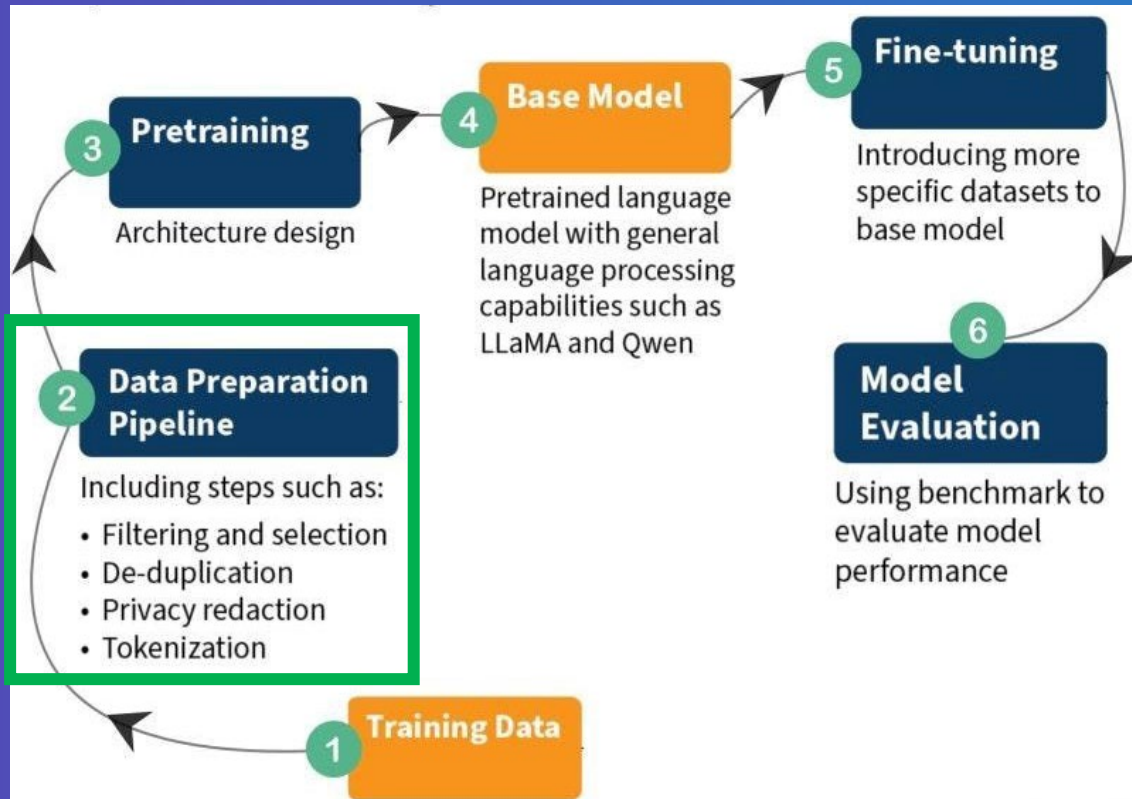


Collaborating on data

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 voor 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

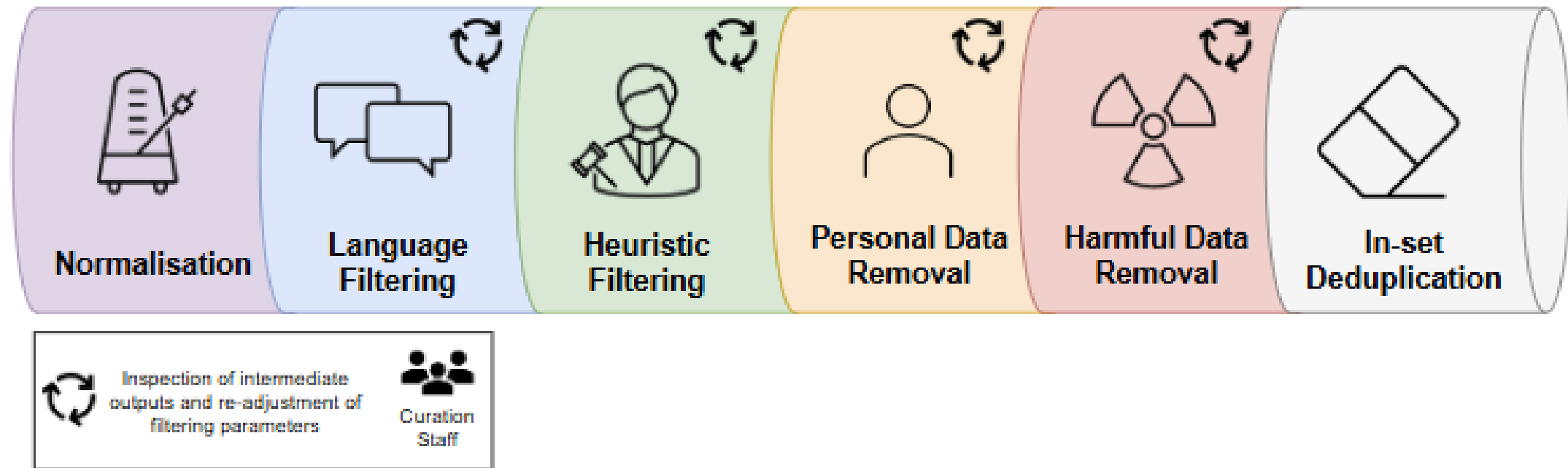
GPT-NL Corpus	NL (B)	EN (B)	Code (B)	Other (B)
High Quality	45	49	231	17
Low-Medium Quality	8.6	159	-	31

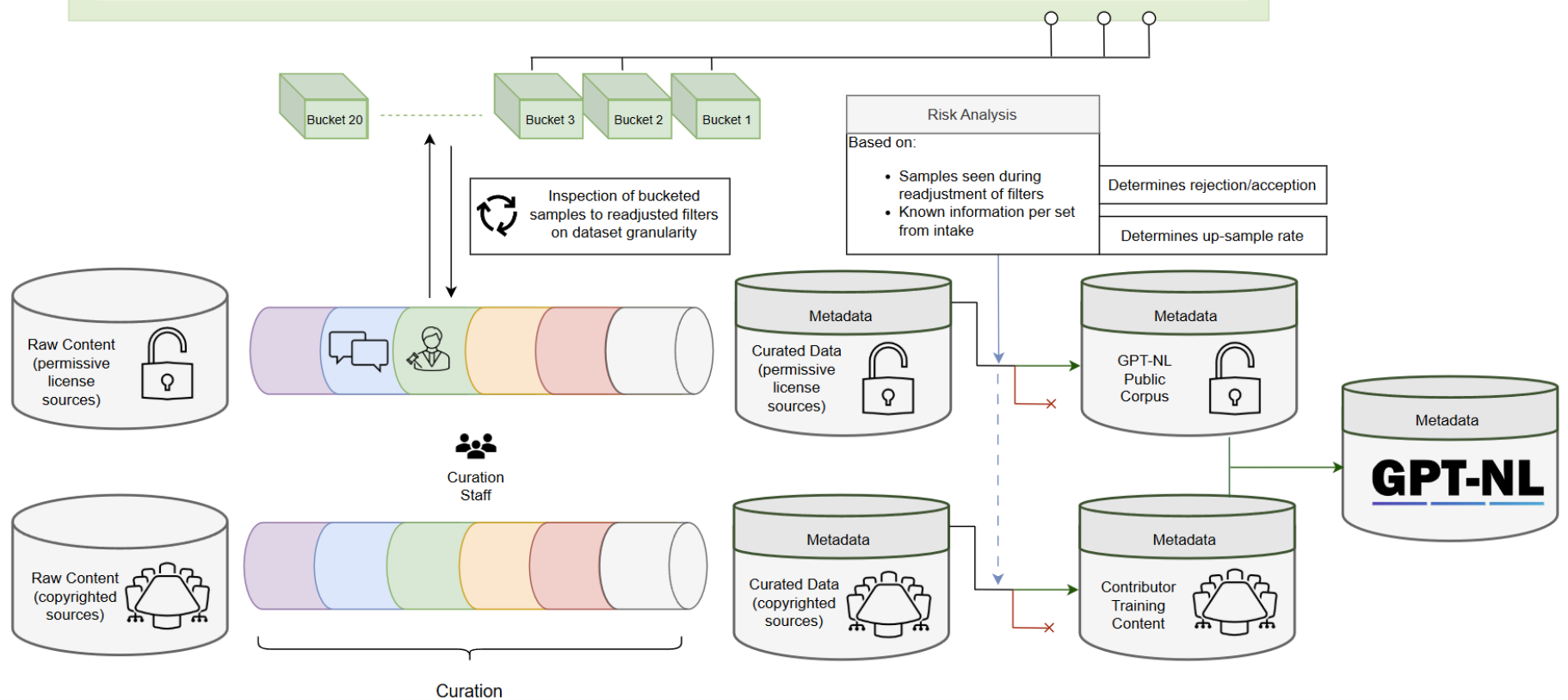
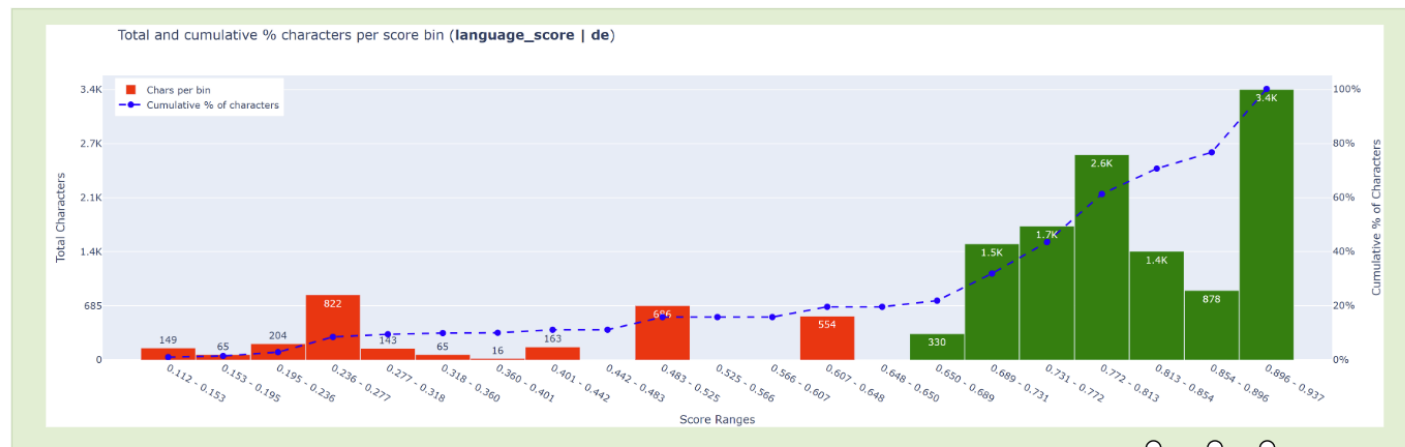


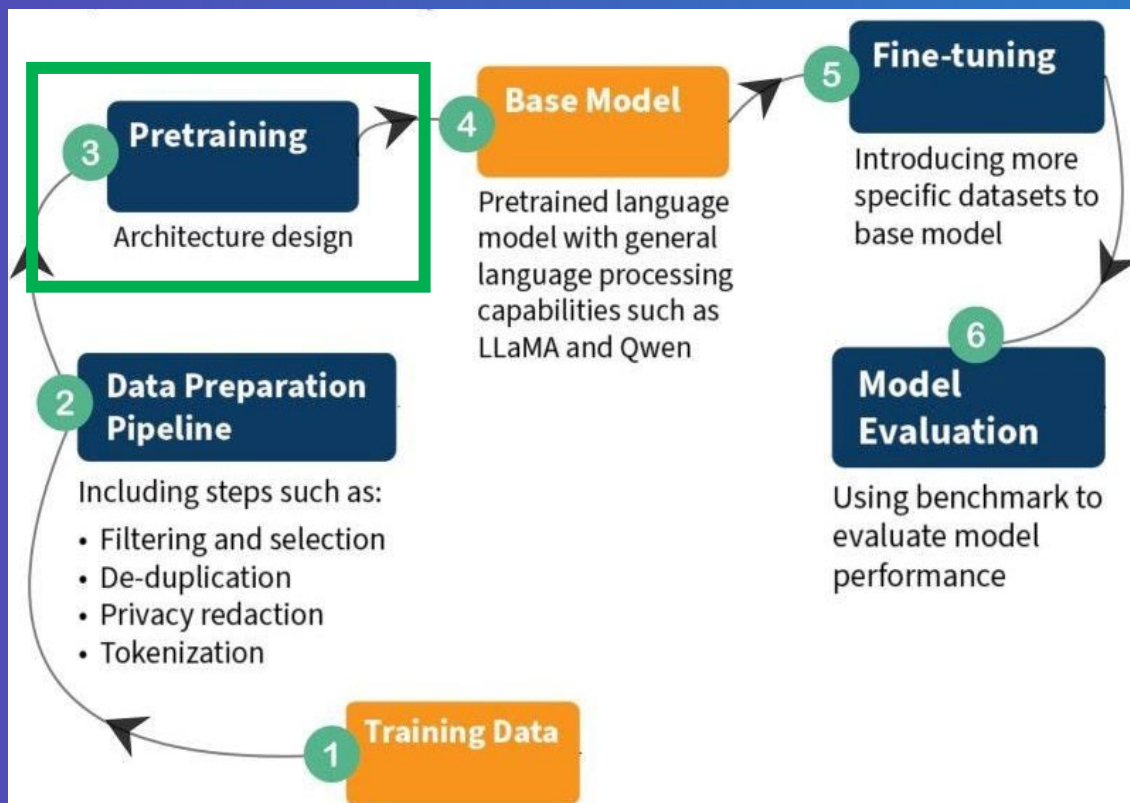
Data Preparation Pipeline

From raw data to curated data

Data Curation Process



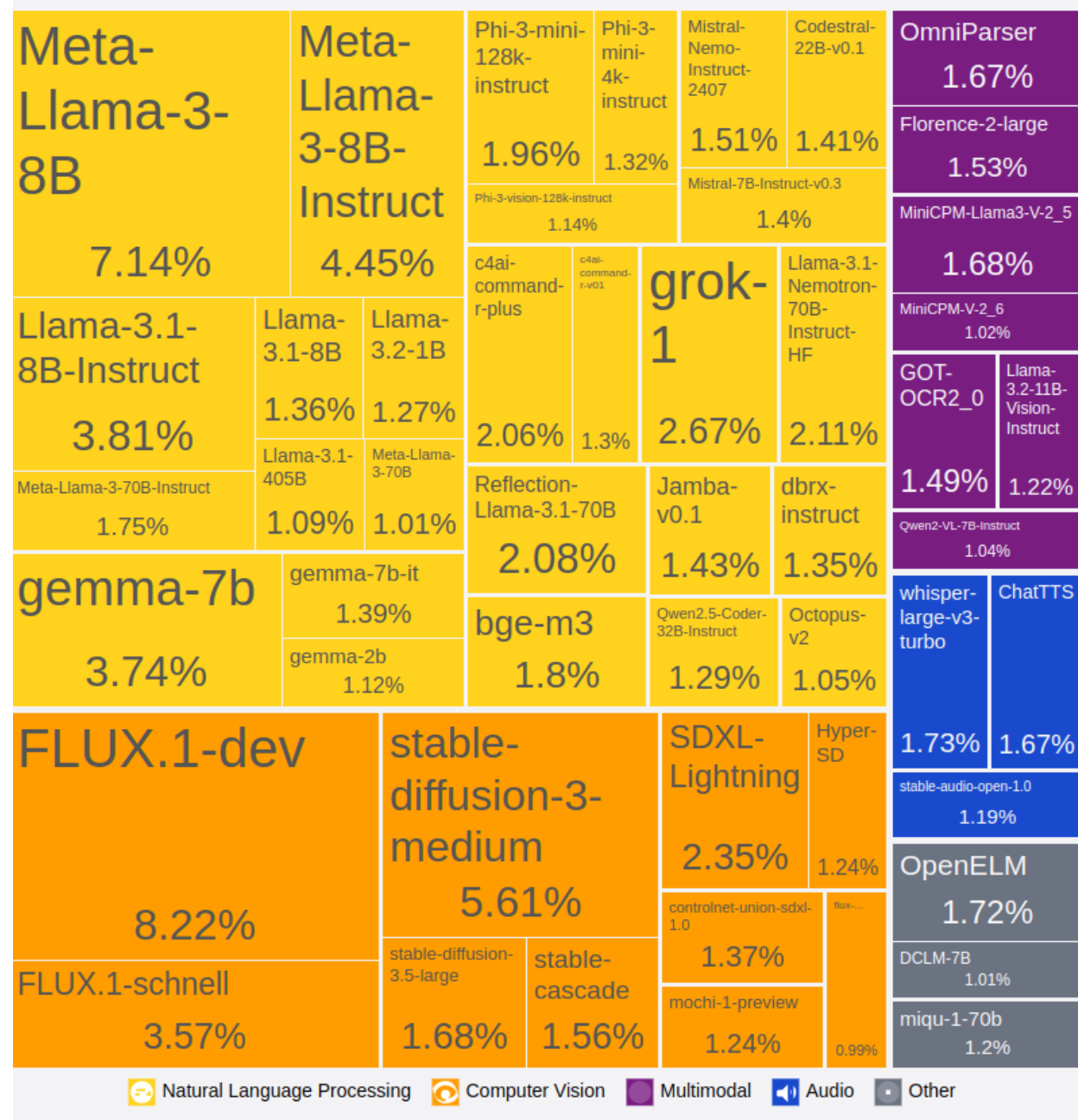




Pretraining

Model design

- Architecture: Llama-3
- Framework: OLMo-core (fully open-source)
- Parameters: 26 billion - GPT-NL 26B

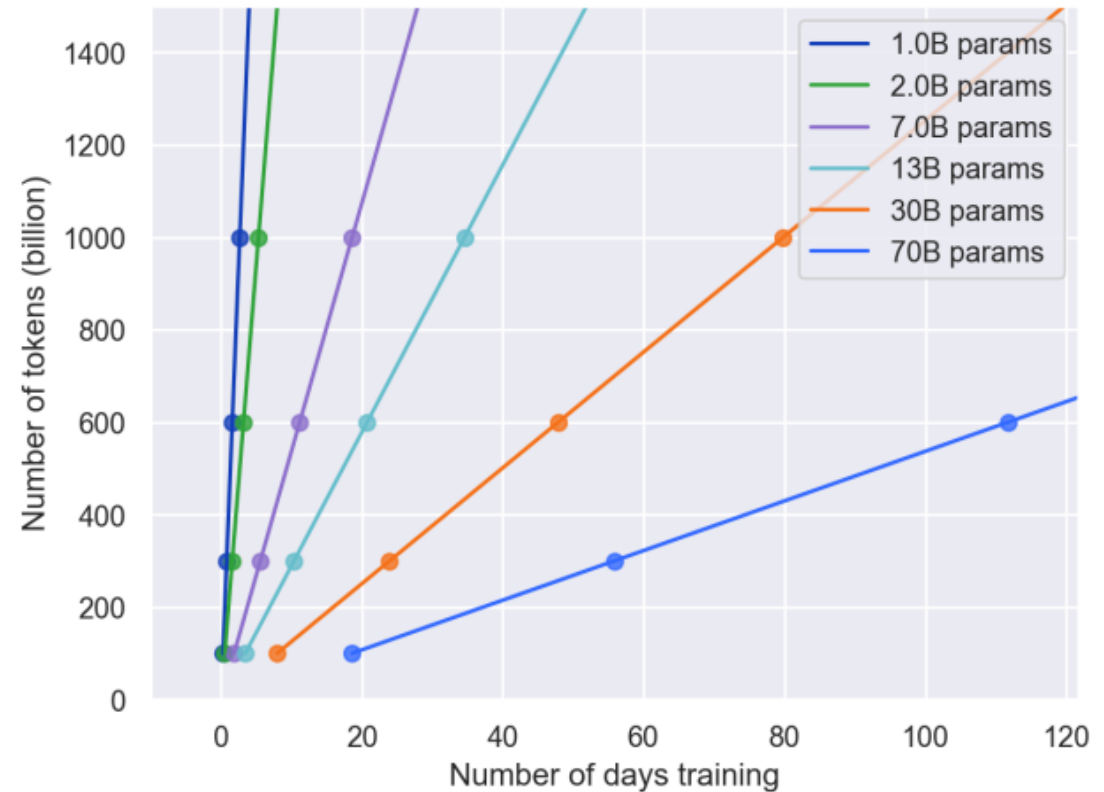


Source: HuggingFace advent calendar 2024 <https://huggingface.co/spaces/huggingface/open-source-ai-year-in-review-2024>

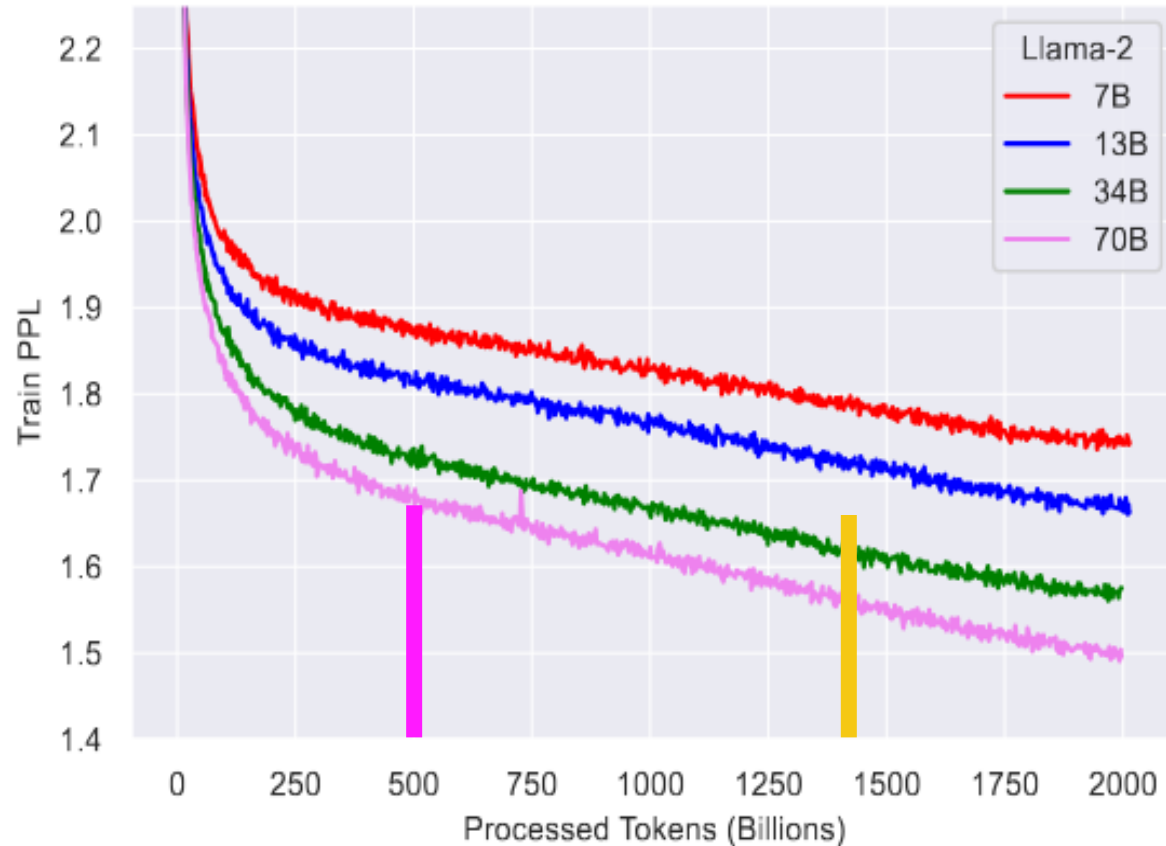
Compute Limitations

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

Estimate T, for different N and D, given a fixed compute (3.7 YFLOPs)
(88 x H100 with 989 TFLOPS, MFU=0.4 & bfloat16)

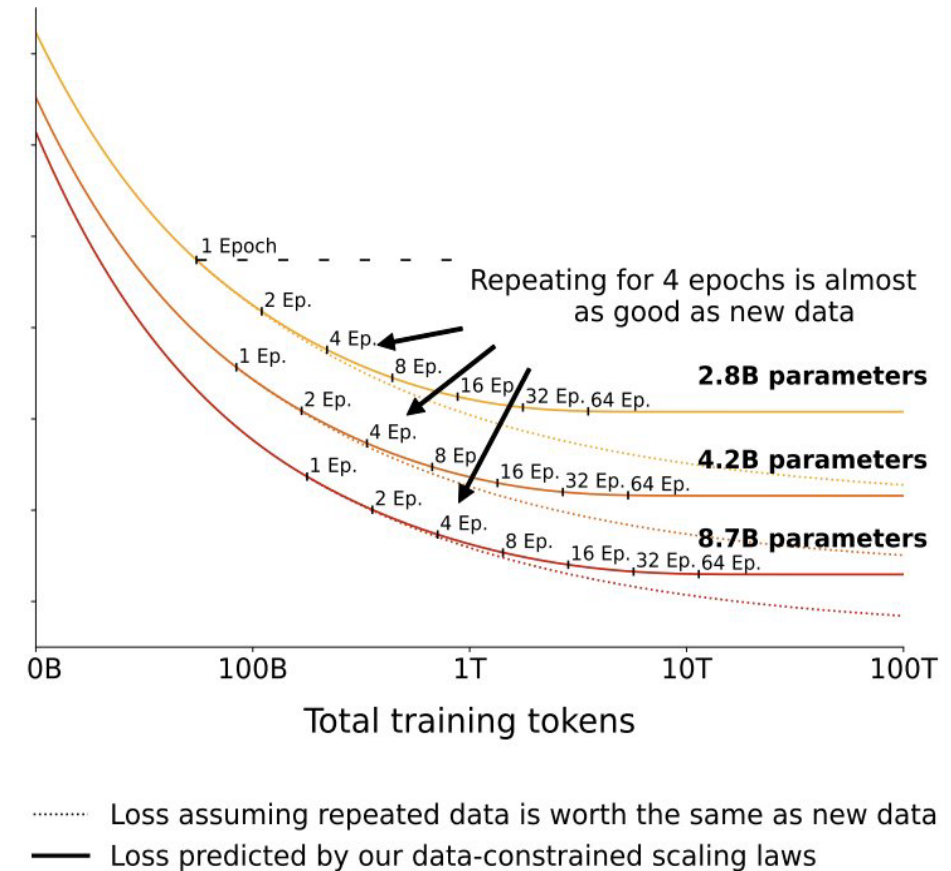


Increasing Model Size versus Oversampling



Source: Reprinted from *Llama 2: Open Foundation and Fine-Tuned Chat Models* by Touvron et al., 2023 (Meta AI Research)

Predicted Loss (Variable training length)



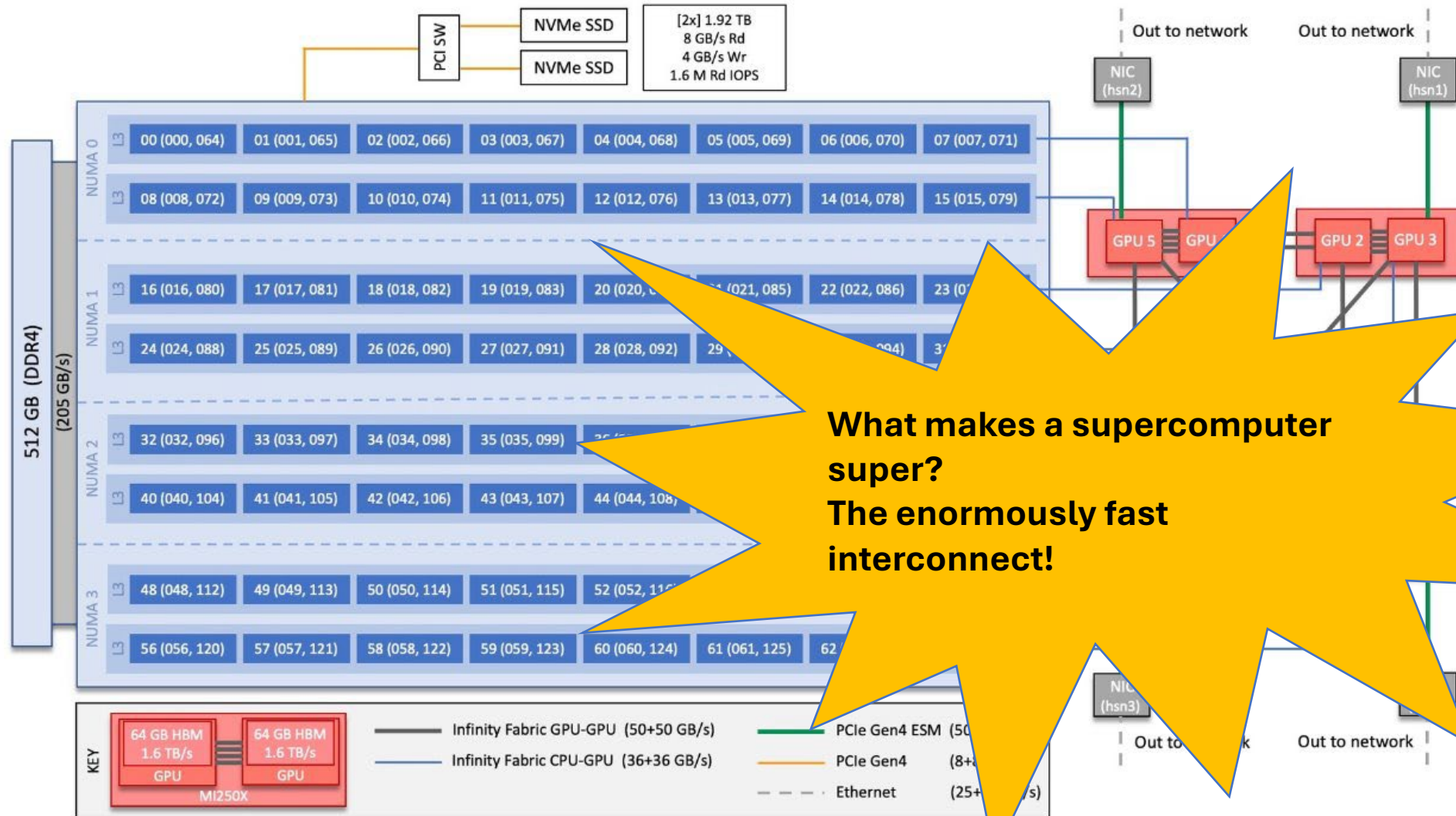
Source: Muennighoff et al., *Scaling Data-Constrained Language Models* (2023)

Snellius – the national supercomputer

- 600+ GPUs
- 200,000+ CPU cores
- NWO subsidized
 - Free for researchers
 - Commercial contracts like GPT-NL
- GPT-NL trains on 88 H100 GPUs

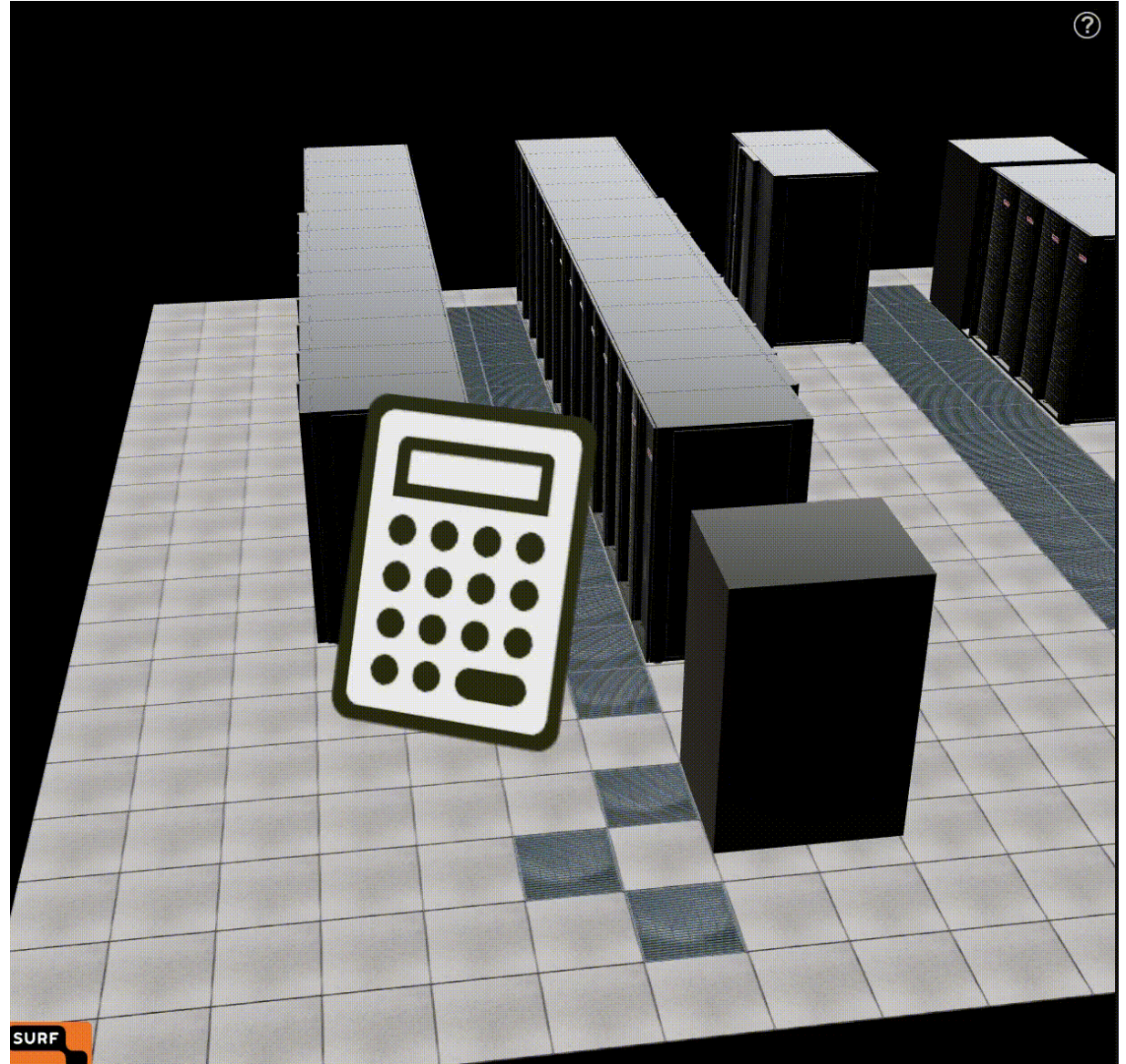


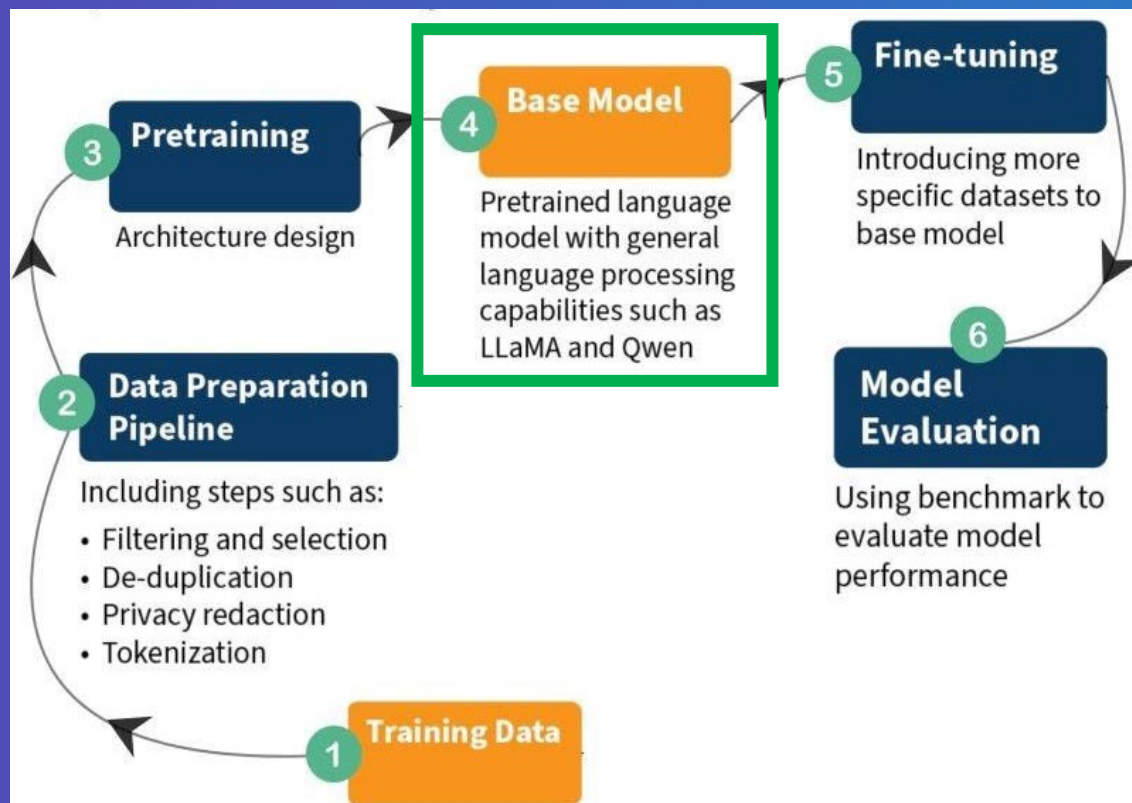
Snellius – the national supercomputer



Snellius

- Training an LLM does not fit within a single GPU
- Shard the model over GPU nodes
- Non-trivial task!



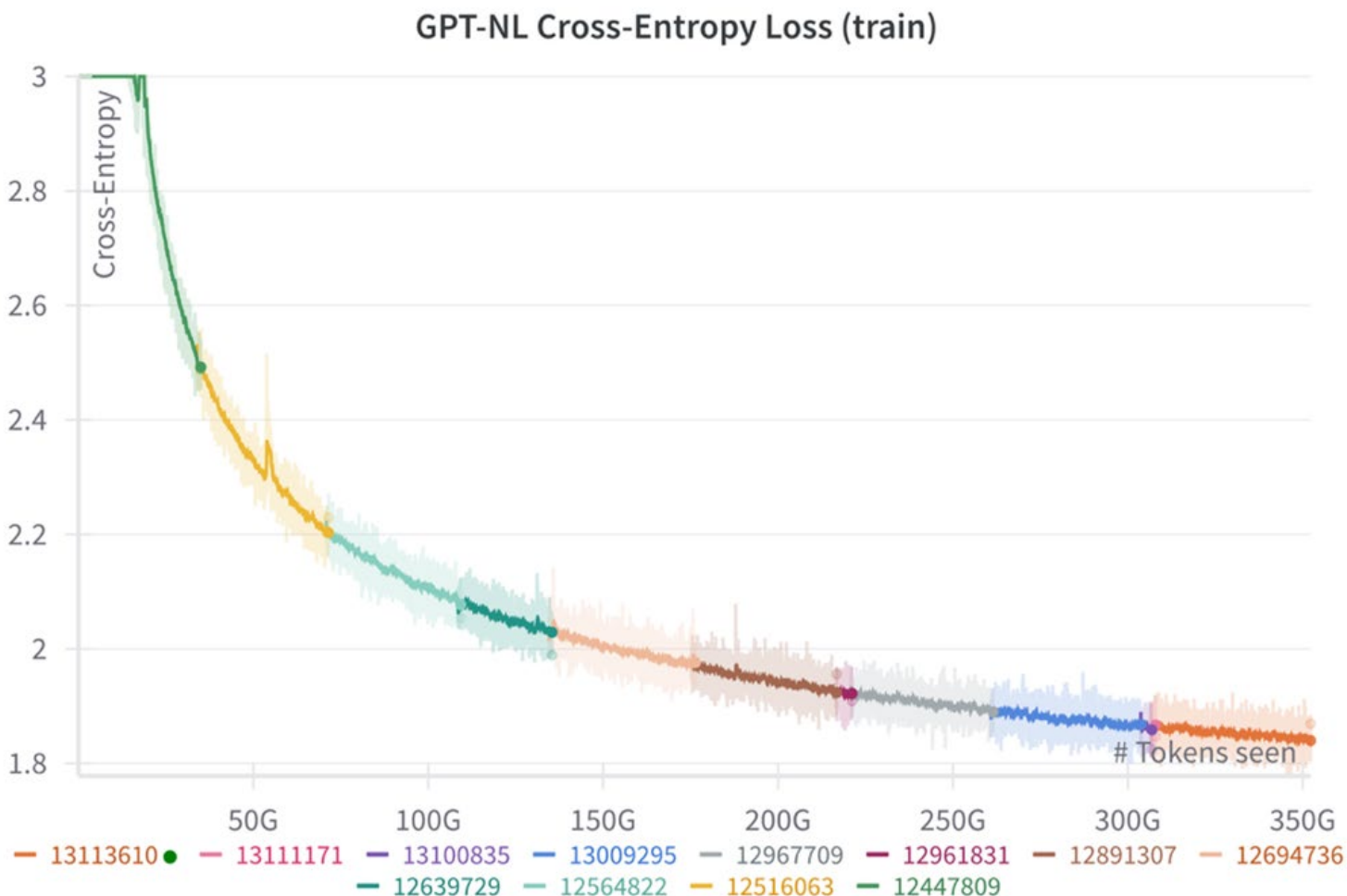


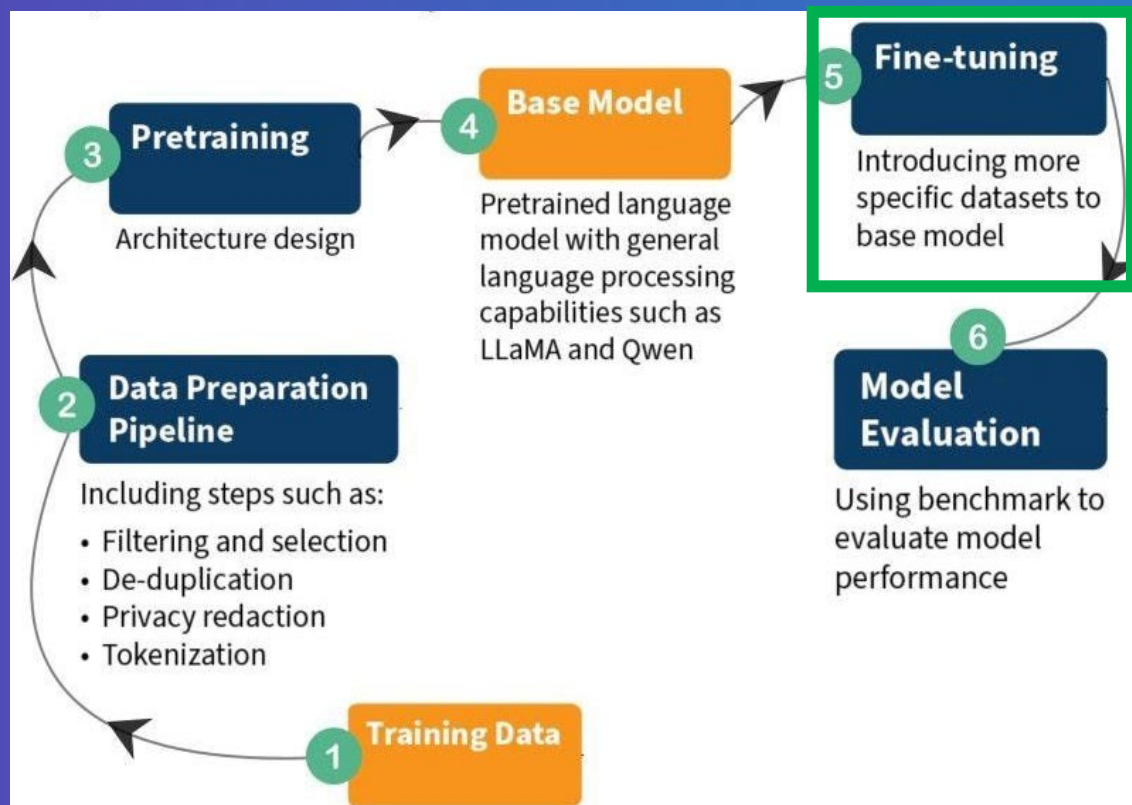
Base Model

Live update of GPT-NL

- Now at 1.2T tokens
- Epoch 3 in progress

```
Device 0 [NVIDIA H100] PCIe GEN 5@16x RX: 12.17 GiB/s TX: 8.308 GiB/s  
GPU 1560MHz MEM 1593MHz TEMP 68°C FAN N/A POW 695 / 700 W  
GPU[|||||100%] MEM[|||||84.633Gi/93.584Gi]  
  
Device 1 [NVIDIA H100] PCIe GEN 5@16x RX: 12.46 GiB/s TX: 1.217 GiB/s  
GPU 1515MHz MEM 1593MHz TEMP 66°C FAN N/A POW 693 / 700 W  
GPU[|||||100%] MEM[|||||84.637Gi/93.584Gi]  
  
Device 2 [NVIDIA H100] PCIe GEN 5@16x RX: 10.69 GiB/s TX: 0.992 GiB/s  
GPU 1425MHz MEM 1593MHz TEMP 68°C FAN N/A POW 699 / 700 W  
GPU[|||||100%] MEM[|||||84.622Gi/93.584Gi]  
  
Device 3 [NVIDIA H100] PCIe GEN 5@16x RX: 8.104 GiB/s TX: 702.7 MiB/s  
GPU 1425MHz MEM 1593MHz TEMP 69°C FAN N/A POW 698 / 700 W  
GPU[|||||100%] MEM[|||||84.610Gi/93.584Gi]
```





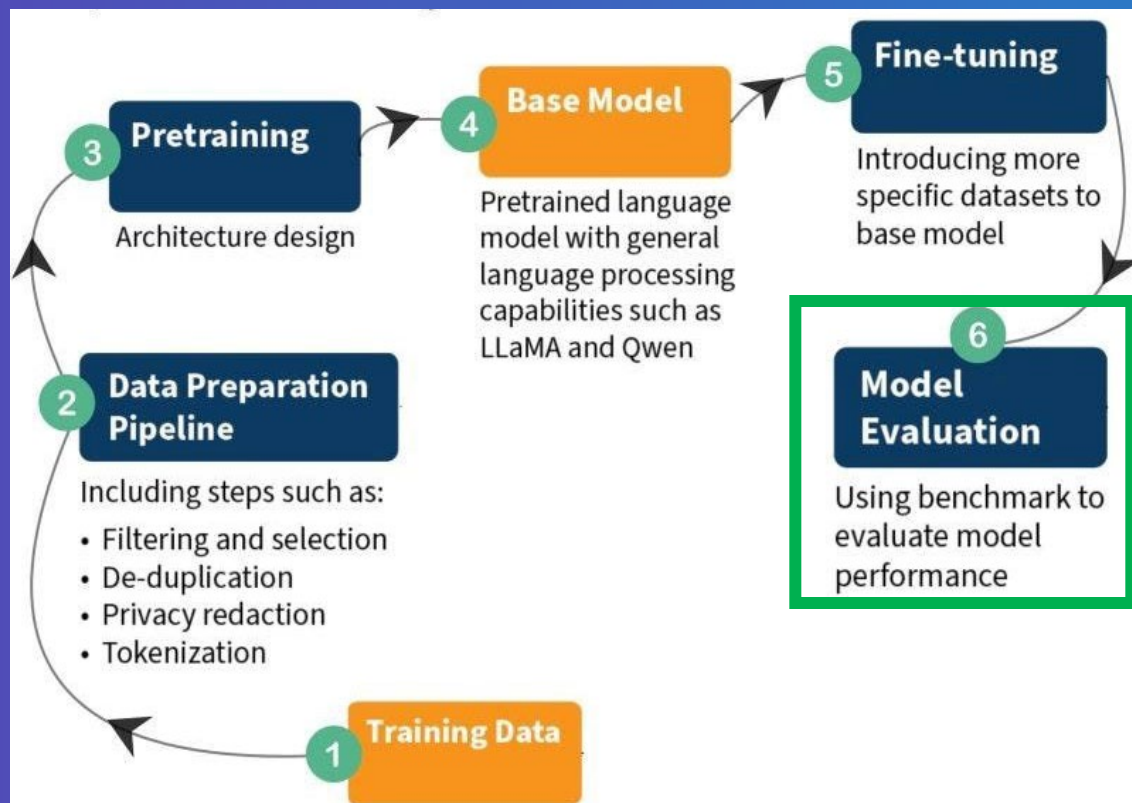
Fine-tuning

Pretraining vs. Finetuning

Aspect	Pretraining	Finetuning
Purpose	Build general language ability	Adapt to tasks & align with users
Data	Massive, diverse, noisy	Small, curated, high-quality
Cost	Very compute- & time-intensive	Relatively cheap & fast
Output	General-purpose base model	Task- or domain-specialized model

Finetuning data

#	Task	Created
3000	Closed question (based on context)	
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	



Evaluation

Evaluation

- Three use-cases
 - Summarization
 - Simplification
 - Retrieval-Augmented Generation (RAG)
- Lack of authentic Dutch benchmarks
 - No generative AI generated content

	Model	Type	Rank ▲	Parameters	Vocabulary	Context
1	meta-llama/Llama-3.1-405B	🧠	1.27	406B	128K	131K
2	deepcogito/cogito-v1-preview-qwen-32B	📝	1.42	33B	152K	131K
3	gpt-5-2025-08-07 (zero-shot, val)	😞	1.43	?	100K	272K
4	meta-llama/Llama-3.1-405B-Instruct-FP8	📝	1.43	406B	128K	131K
5	gpt-5-2025-08-07#high (zero-shot, val)	😞	1.43	?	100K	272K
6	o3-2025-04-16 (zero-shot, val)	😞	1.45	?	?	200K
7	mistralai/Mistral-Small-3.1-24B-Instruct-2503	📝	1.46	24B	131K	?
8	gemini/gemini-2.5-pro (zero-shot, val)	😞	1.46	?	256K	1M
9	deepcogito/cogito-v2-preview-llama-70B	📝	1.47	71B	128K	131K
10	gpt-5-2025-08-07#minimal (zero-shot, val)	😞	1.47	?	100K	272K

Source: EuroEval <https://euroeval.com/leaderboards/Monolingual/dutch/>

GPT-NL benchmarks

- Authentic Dutch datasets
- Factual knowledge
 - Dutch integration exam
 - Integration High school exams in collaboration with UvA
- Simplification
 - Integration DuidelijkeTaal in collaboration with Instituut voor de Nederlandse Taal
- Bias benchmark
 - MBBQ integration in collaboration with UvA
- Common **research** benchmarks
 - Good proxy but not ideal for in production



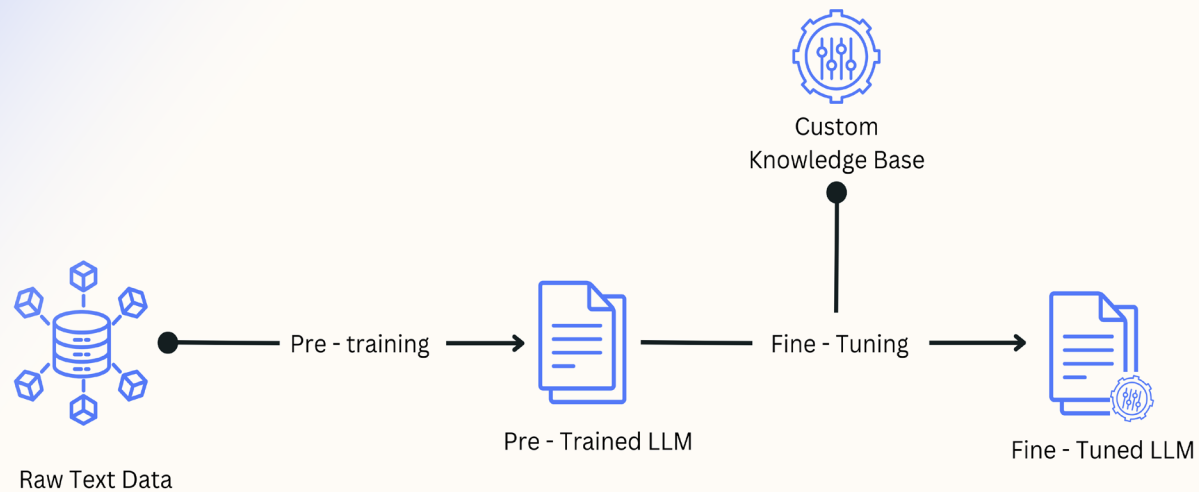
The robust European language model benchmark.

Source: EuroEval <https://euroeval.com/>

Finetuning

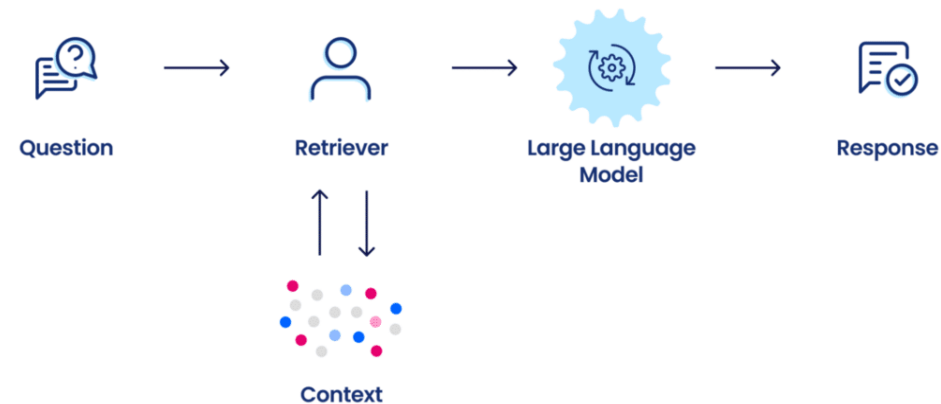
vs.

RAG



Source: Figure from TagX Data, "What is Supervised Fine-Tuning?"

Retrieval Augmented Generation



Source: Figure from Snorkel AI Blog — Hoang Tran, "Which is better, retrieval augmentation (RAG) or fine-tuning? Both." (2023, updated 2024)

Evaluation

- What about in production?
 - Traditional NLP metrics
 - LLM-as-a-judge
 - RAG benchmark

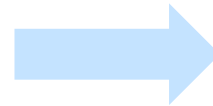
Test case

Input prompt – query to the model/system

Actual Output – the LLM's response

Expected Output (optional) - ground truth output

Context (optional) – retrieval or background information for RAG



Metrics

Relevancy

Toxicity

Faithfulness

Hallucination

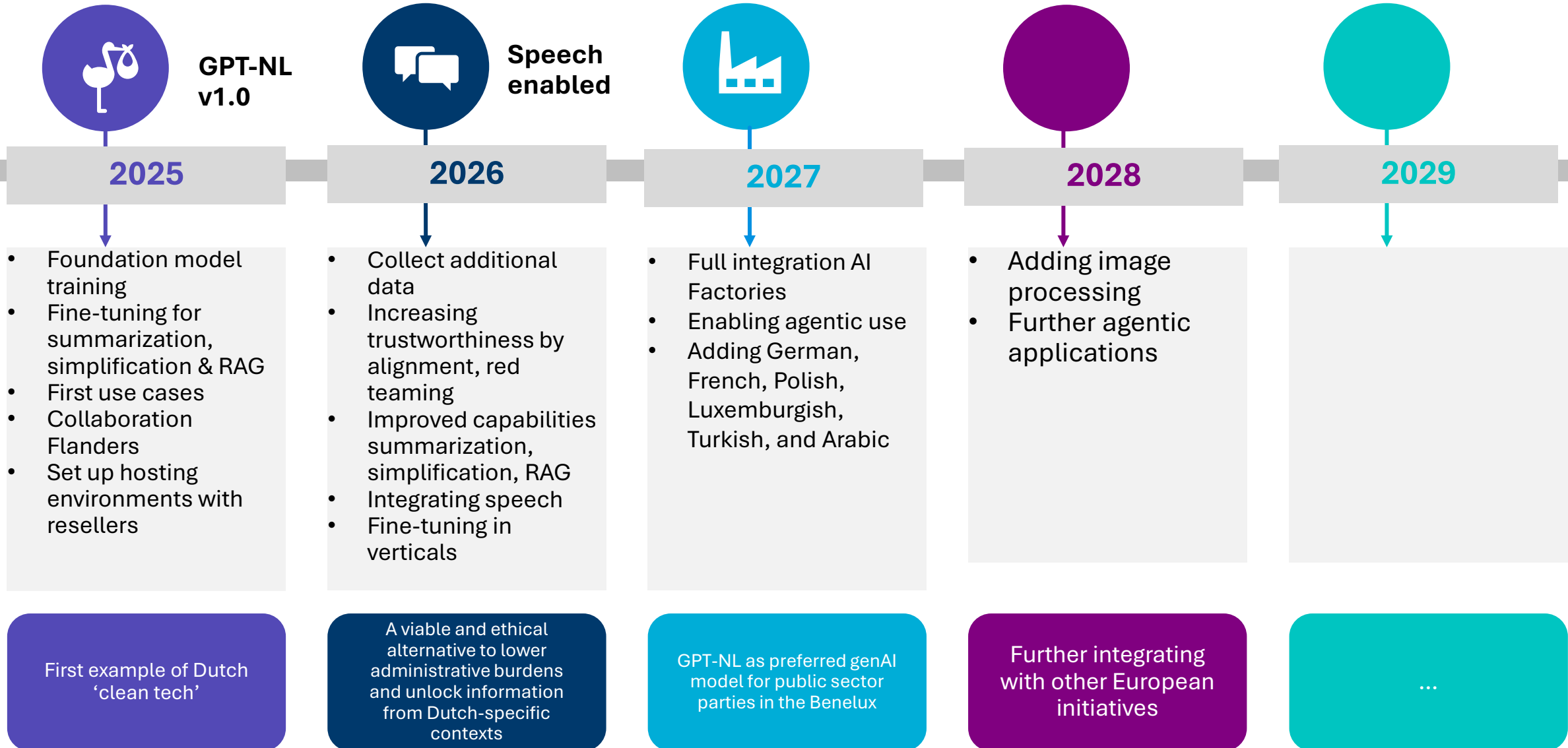


DeepEval.

\$ the open-source LLM evaluation framework

Source: DeepEval <https://deepeval.com/>

Dream big



More Information



jesse.vanoort@tno.nl



Website GPT-NL (never scan
QR codes you don't trust)

www.gpt-nl.nl

GPT-NL

TNO innovation
for life



Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

