GPT-NL

FACILITEIT VOOR EEN SOEVEREIN NEDERLANDS TAALMODEL

CLIN 2024

GPT-NL team

Dominique Blok & Erik de Graaf

TNO innovation for life

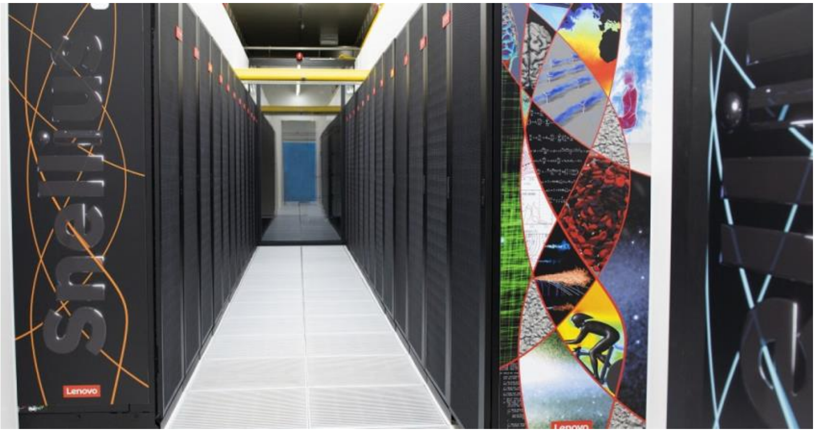
Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

SURF

Consortium

TNO innovation for life

SURF



Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

Why GPT-NL?

2022: ChatGPT

ChatGPT: New AI chatbot has everyone talking to it

BBC

FINANCIAL TIMES

Opinion **Artificial intelligence**

ChatGPT is fluent, clever and dangerously creative

nrc›

Als de computer beter wordt met taal dan wij

The Verge

ChatGPT proves AI is finally mainstream – and things are only going to get weirder

TechScape: Meet ChatGPT, the viral AI tool that may be a vision of our weird tech future

the Guardian

Is Chat GPT the world's first truly useful chatbot?

THE TIMES

2022: ChatGPT

The New York Times

Will ChatGPT Make Me Irrelevant?

CNN

Is no career safe anymore?

Is AI coming for your job? ChatGPT renews fears

abc NEWS

nature

AI bot ChatGPT writes smart essays – should professors worry?

How Generative AI Will Change All Knowledge Work

TIME

the Guardian

What is AI chatbot phenomenon ChatGPT and could it replace humans?

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

Criticism of LLMs

OpenAI's hunger for data is coming back to bite it

MIT Technology Review

Transparency is sorely lacking amid growing AI interest

ZDNET

Evening Standard.

Meta's use of user data to train its AI violates GDPR, privacy group says

THE NEW STATESMAN

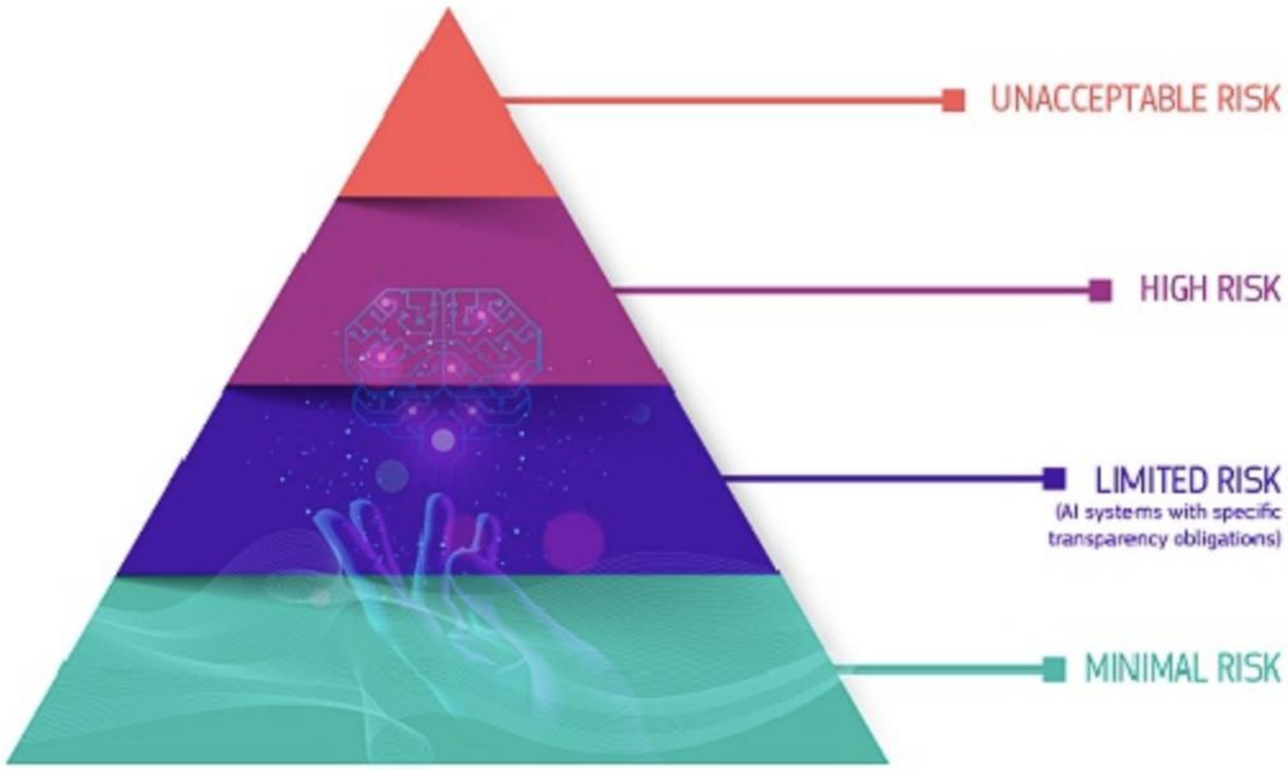
ChatGPT proves that AI still has a racism problem

Former OpenAI employees say whistleblower protection on AI safety is not enough

The Verge

EU AI Act

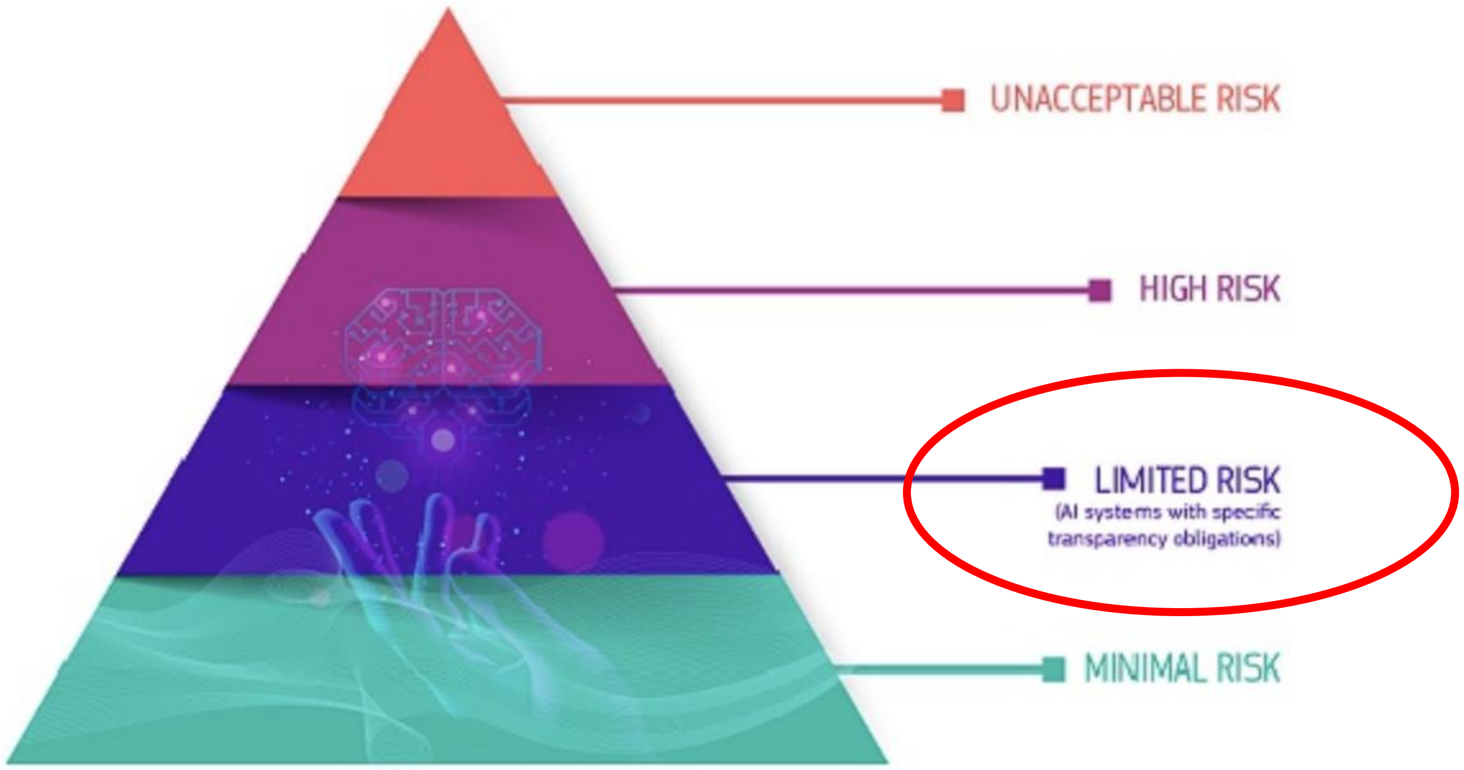
The Regulatory Framework defines 4 levels of risk for AI systems:



- Officially entered into force on August 1st, 2024
- Risk-based approach

EU AI Act

The Regulatory Framework defines 4 levels of risk for AI systems:

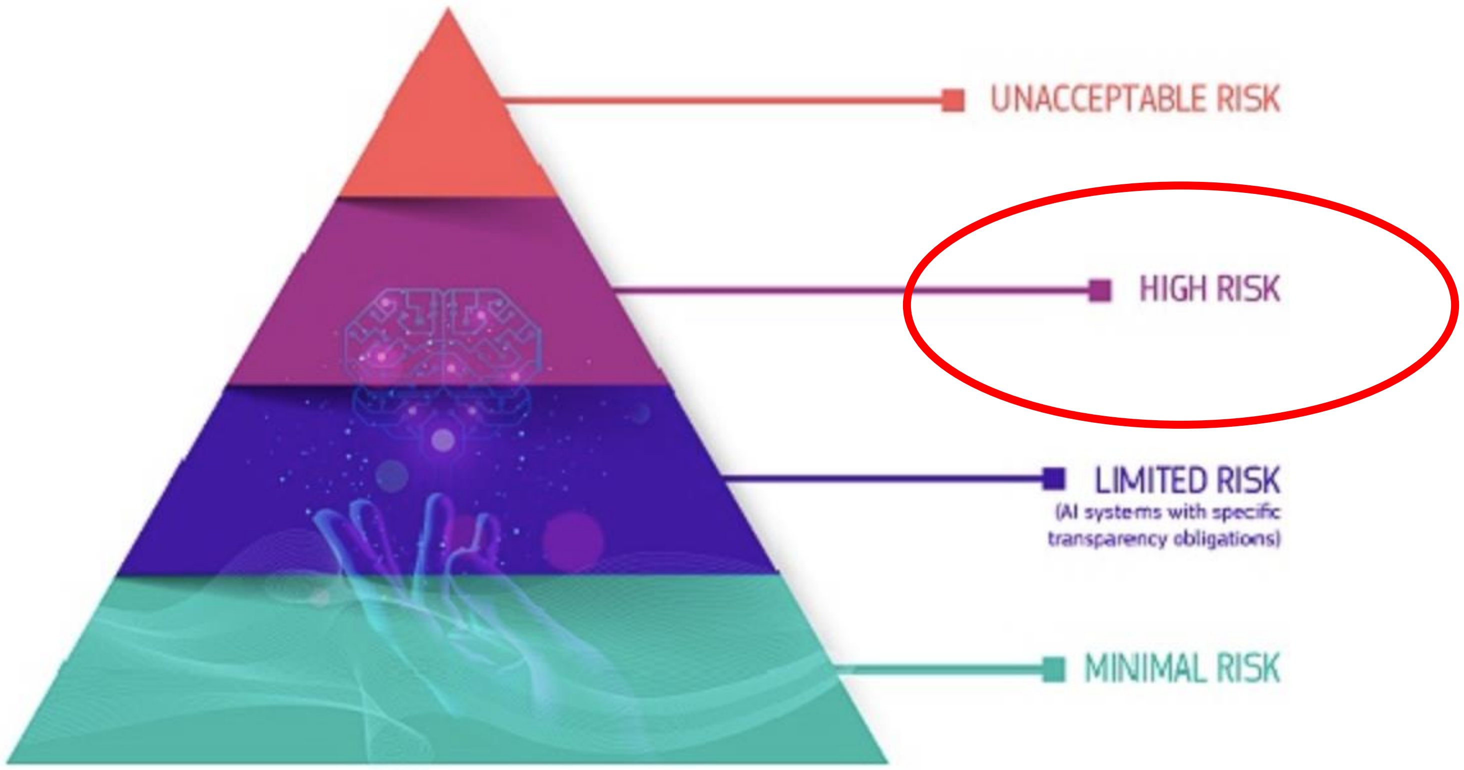


- Allowed as long as one is transparent about use of AI
- Examples: chatbots, AI generated text

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

EU AI Act

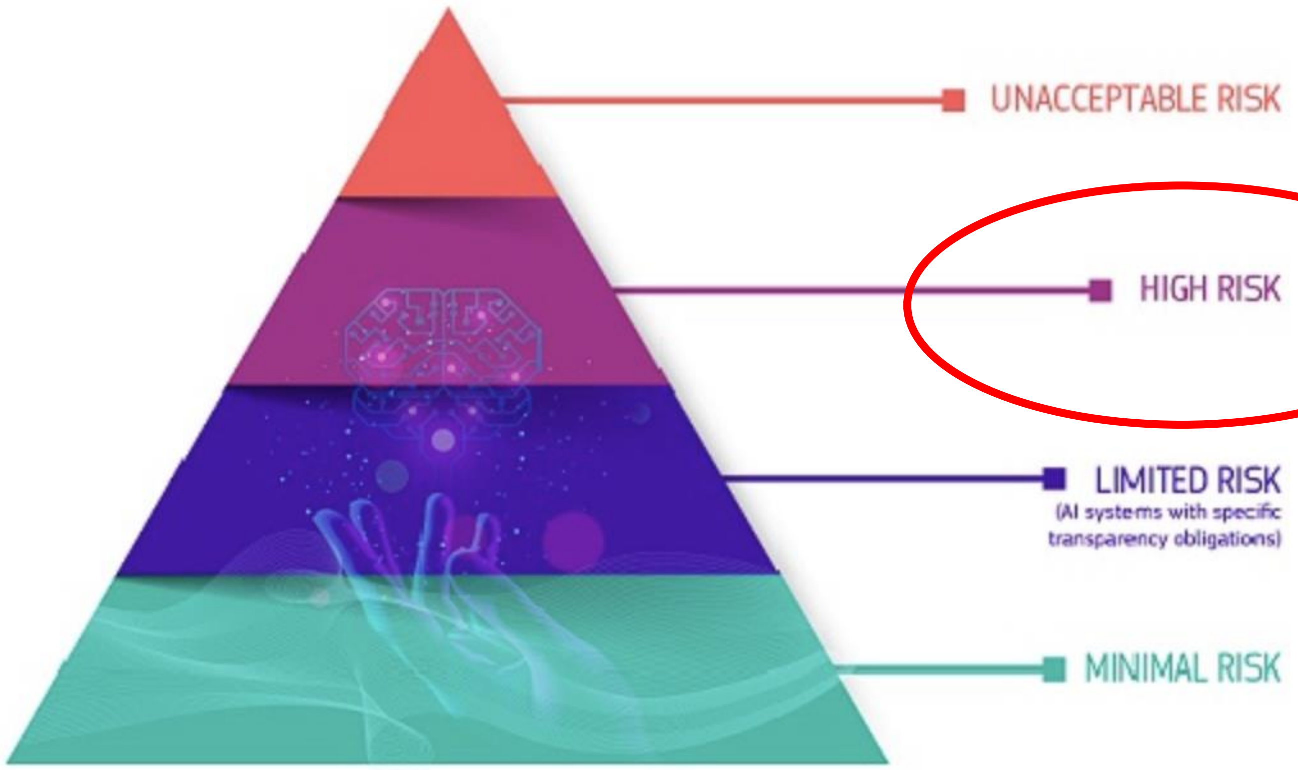
The Regulatory Framework defines 4 levels of risk for AI systems:



- Subject to strict obligations
- Examples: critical infrastructure, education, essential public services such as healthcare, law enforcement, border management, justice and democratic processes

EU AI Act

The Regulatory Framework defines 4 levels of risk for AI systems:



Obligations:

- adequate risk assessment and mitigation systems
- high quality of the datasets feeding the system to minimise risks and discriminatory outcomes
- logging of activity to ensure traceability of results
- detailed documentation providing all information necessary on the system and its purpose for authorities to assess its compliance
- clear and adequate information to the deployer
- appropriate human oversight measures to minimise risk
- high level of robustness, security and accuracy

Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

Legal requirements

- AI Act:
  - Transparency
  - Robustness
  - Safety
- GDPR (General Data Protection Regulation): privacy
- Intellectual Property Law: prohibits taking IP without permission

As compliant as possible

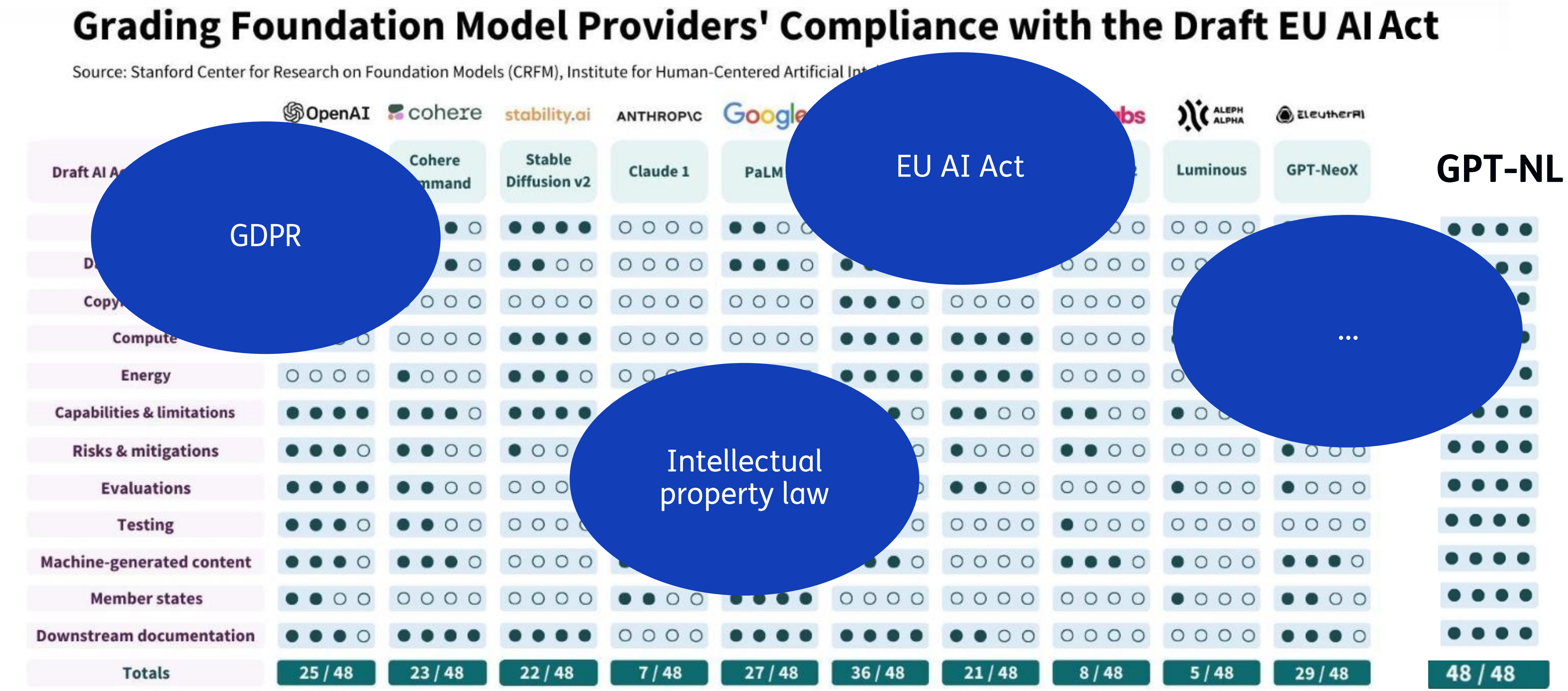
Grading Foundation Model Providers' Compliance with the Draft EU AI

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

https://crfm.stanford.edu/2023/06/15/eu-ai-act.html

As compliant as possible



Why a Dutch LLM from scratch?

Besides **compliance with legislation**…

- Many of the current language models are trained on datasets that contain **no or very little Dutch data**
- **European values around bias and inclusivity** are insufficiently guaranteed in current solutions
- **Digital sovereignty** of European language and speech technology, no dependence on foreign multinationals
- Our **commitments** to a better AI ecosystem: https://gpt-nl.nl/commitments/

de Volkskrant

NIEUWS

Nederland ontwikkelt antwoord op ChatGPT: AI-taalmodel GPT-NL

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.

Chinese organisations launched 79 AI large language models since 2020, report says

Große KI-Modelle

FÜR DEUTSCHLAND

Machbarkeitsstudie 2023

LEAM:AI

KI BUNDESVERBAND

◆ WSJ NEWS EXCLUSIVE

Europe to ChatGPT: Disclose Your Sources

Proposed legislation requires developers to list copyright material used in generative AI tools

Why do we need a large GPT for Swedish?

What are the advantages of building a large language model for Swedish, and what should we look out for?

Magnus Sahlgren · Follow

Published in AI Sweden · 6 min read · Jul 14, 2022

What is GPT-NL?

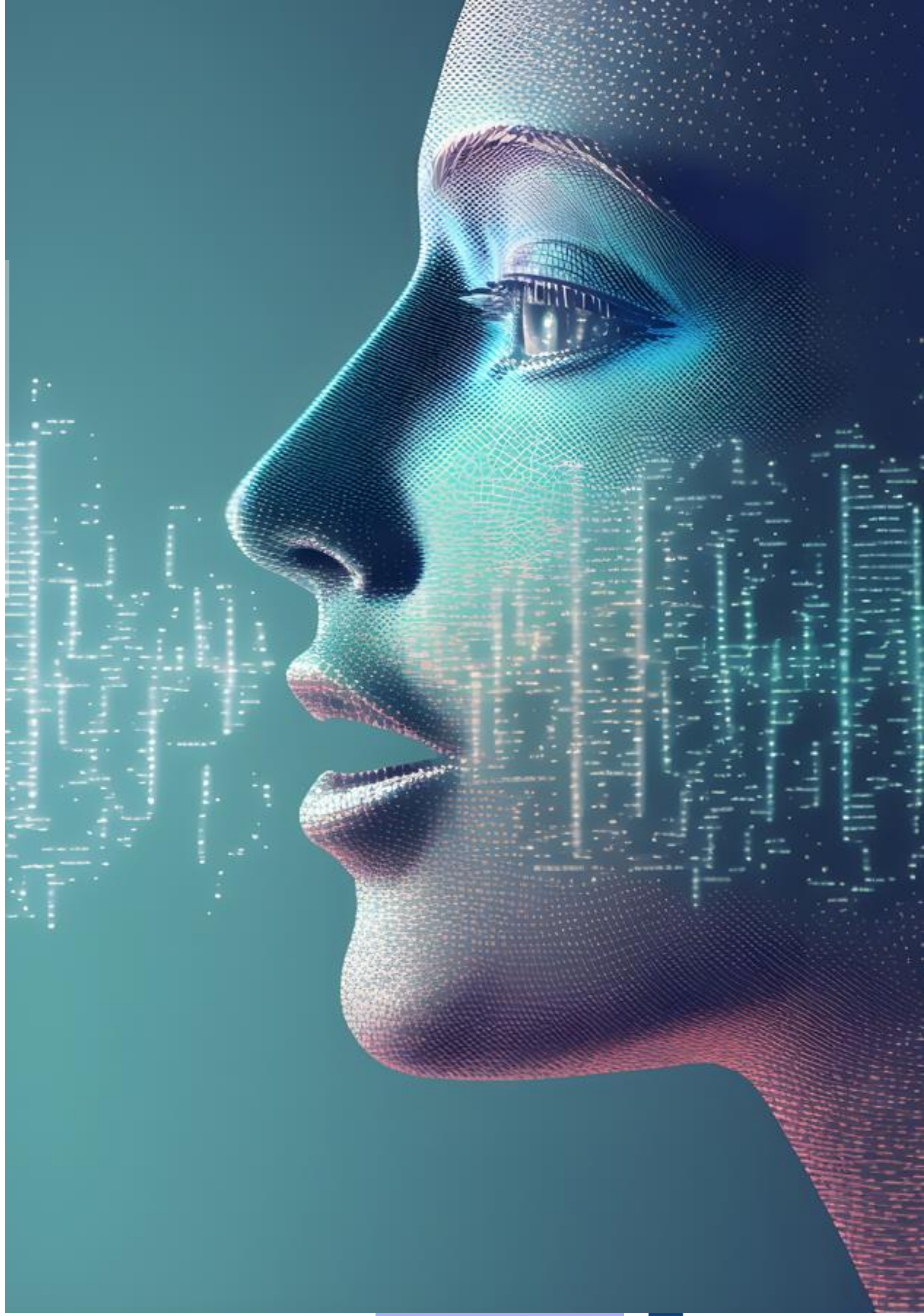
What?

We will build our own Dutch-English (50%-50%) language models from scratch

using data that we are allowed to use,
with privacy information removed,
with full transparency in our choices

Where we strive to be as transparent and compliant as possible

GPT-NL commitments: gpt-nl.nl/commitments

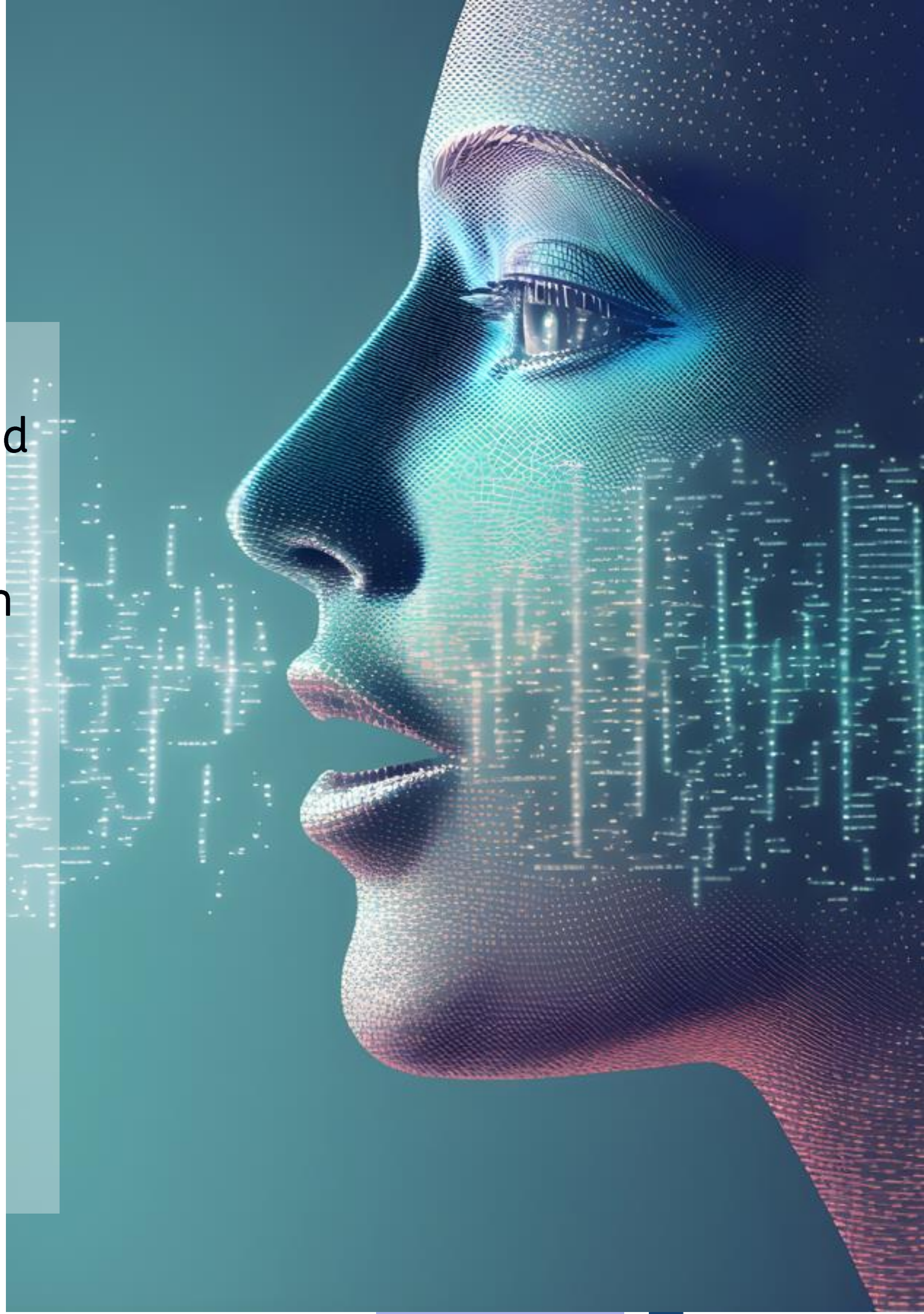


What?

Deliverables:

- We will release the **foundation** model, an **instruct fine-tune**, and a **feedback fine-tune**
- Model architecture still depends on what is state-of-the-art when training starts; around 70B parameters; probably based on Llama architecture

FOUNDATION MODEL	INSTRUCT MODEL	
RAW TEXT DATA	INSTRUCTIONS	FEEDBACK



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

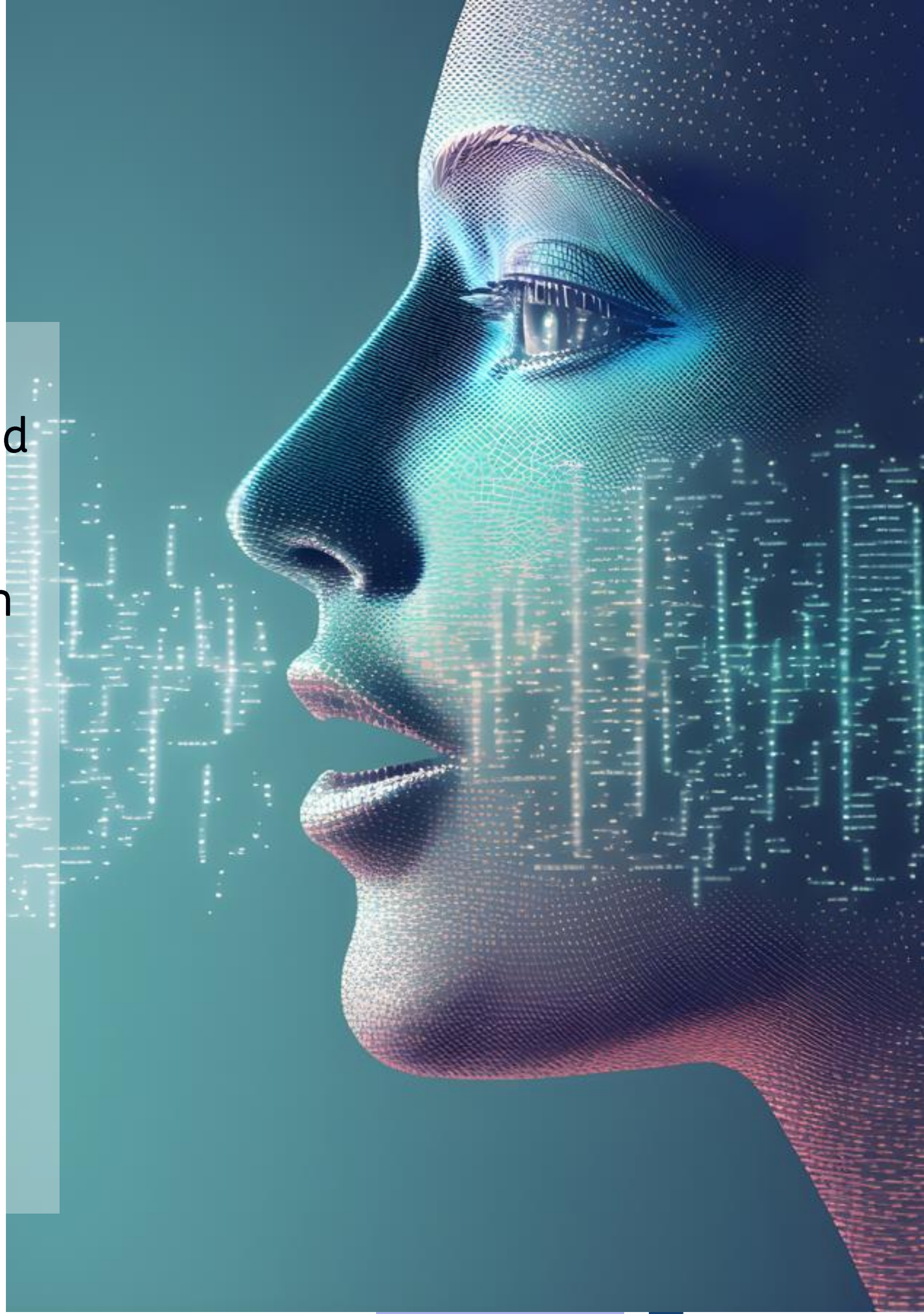
What?

Deliverables:

- We will release the **foundation** model, an **instruct fine-tune**, and a **feedback fine-tune**
- Model architecture still depends on what is state-of-the-art when training starts; around 70B parameters; probably based on Llama architecture

Focus on three main capabilities:

1. **Summarisation**
2. **Simplification**
3. **Retrieval-Augmented Generation (RAG)**
4. **Chat**
5. **Brainstorming**
6. **Open/closed QA**

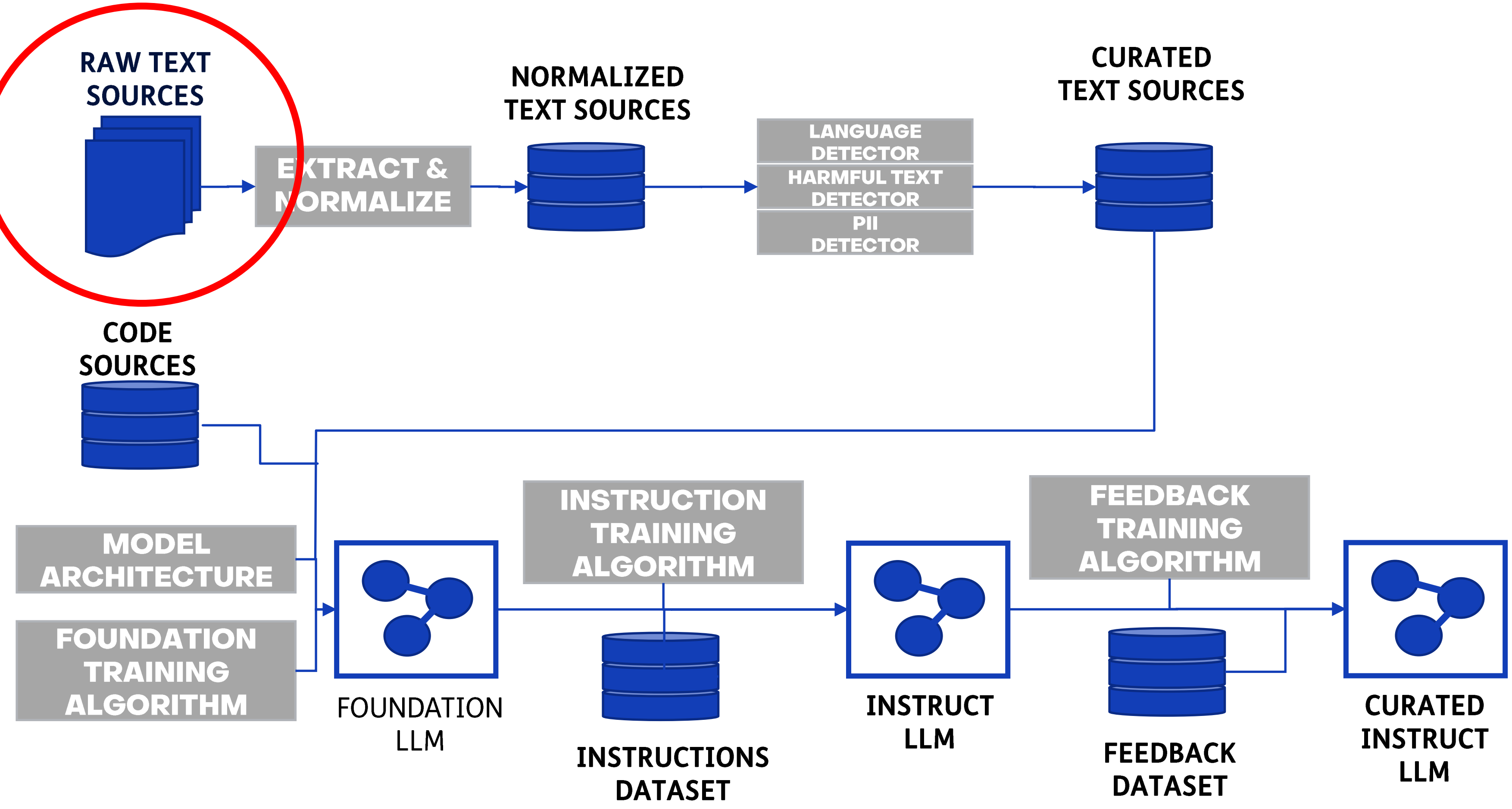


SURF TNO innovation for life Nederlands Forensisch Instituut Ministerie van Justitie en Veiligheid

Data acquisition

GPT-NL

The road to GPT-NL



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

Data acquisition

- In progress
- Aim: at least 300B tokens
- Sources:
  1. Data providers
  2. Permissively licensed high quality data from the web
  3. Synthetic data
  4. Code data (±40%)

Data acquisition

300B tokens in context:

- ± 3 million x the first Harry Potter book
- ± 7.5 x Gigacorpus †
- ± 6 x all Dutch newspapers and magazines

- 2% of Llama 3's training data

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

Data acquisition

Conditions:

- Ethically obtained data:
  - sources with permissive licenses
  - based on agreements with data holders
  - in accordance with IP law

COMPUTABLE

Stichting Brein haalt ai-training-dataset offline

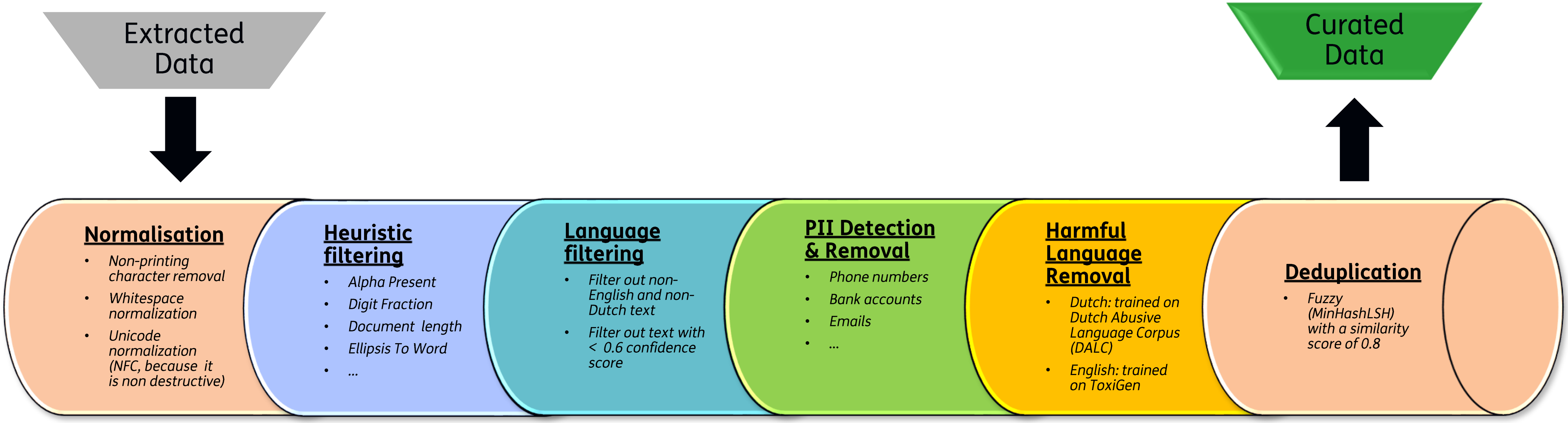
SURF TNO innovation for life Nederlands Forensisch Instituut Ministerie van Justitie en Veiligheid

Data acquisition

Conditions:

- High quality data: no large web scrapes, social media data, etc.
- As much variety in representation as possible:
  - Representation of different groups (different ethnic backgrounds, genders, etc.)
  - Representation of different language varieties and dialects

Data curation



Architecture

- We are training from scratch
- Basing on Llama (3)'s architecture
  - Openly available
  - Great performance

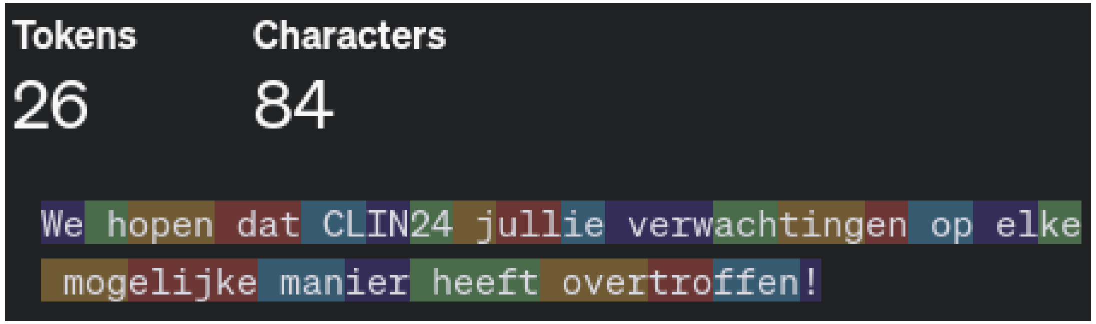
- Final decision to come closer to training
  - Allowing us to adapt to the latest and greatest

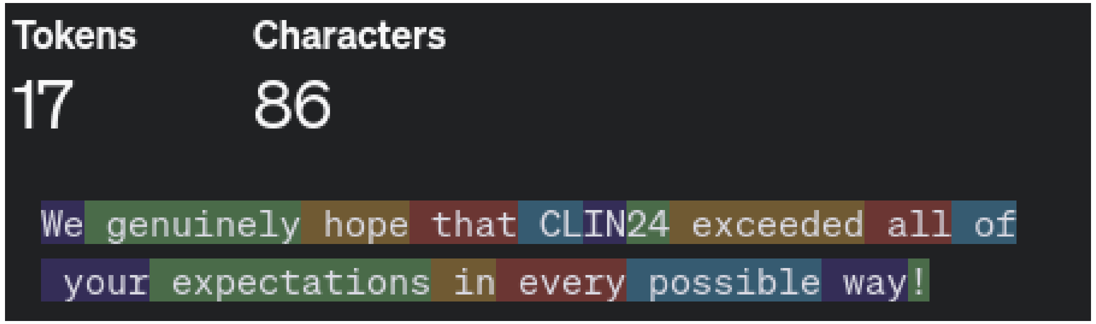
Source: https://github.com/meta-llama/llama3

Tokenizer

- LLMs see tokens rather than letters
- Tokenizers have a vocabulary size (~50k)
- Common tokenizers prioritize English
  - Those tokenizers require more tokens for Dutch
  - More expensive
  - More compute

- We need to train **our own tokenizer**, that fits our dataset





GPT-3.5 & GPT-4 tokenizer sample

https://platform.openai.com/tokenizer

Call to action: Let's make a great Dutch LLM together

What is in it for you?

- Exploitation will go via a license for non-commercial use (free or cheap) and for commercial use (paid).

- You are helping to create a model which takes consideration for privacy, transparency and our common Dutch norms and values.
- The LLM will perform better for your use case if it is trained on similar data.
- You will be financially compensated based on the quantity, quality and diversity of your dataset. The exact calculation for this will follow but 50% of revenue from commercial licenses will go back to data contributors.
- We can offer help with curating data. You get to keep ownership over this curated data.

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

How can you become a contributor?

https://gpt-nl.nl/publish/pages/5387/gpt-nl_data_acquisition_pipeline_en_.pdf

Connect to data providers in your organisation/network as soon as you can and ask to fill in: https://survey.tno.nl/vdwbspltqm?l=nl

If possible work together as data contributors.

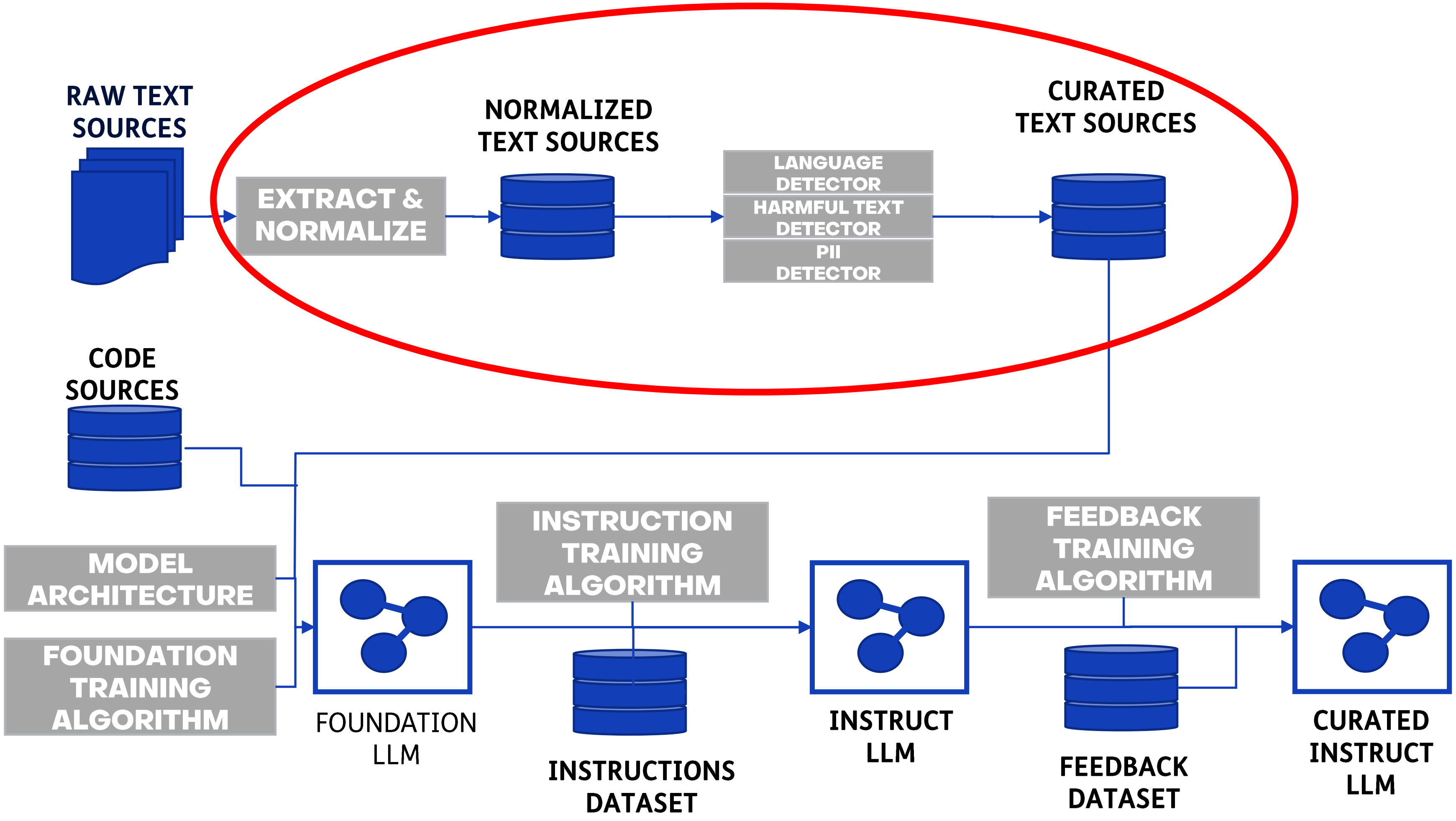
We will come back as soon as possible after we have assessed the survey with a plan of action.

You will also get on our newsletter mailing list.

Data curation

GPT-NL

The road to GPT-NL



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

Data curation: Personally Identifiable Information (PII)

Current method: combination of regex and NER to find:

Regex	NER
E-mail addresses	Names
Addresses	
Passport numbers	
IP addresses	
Crypto addresses	
Phone numbers	
Bank account numbers	
Etc.	

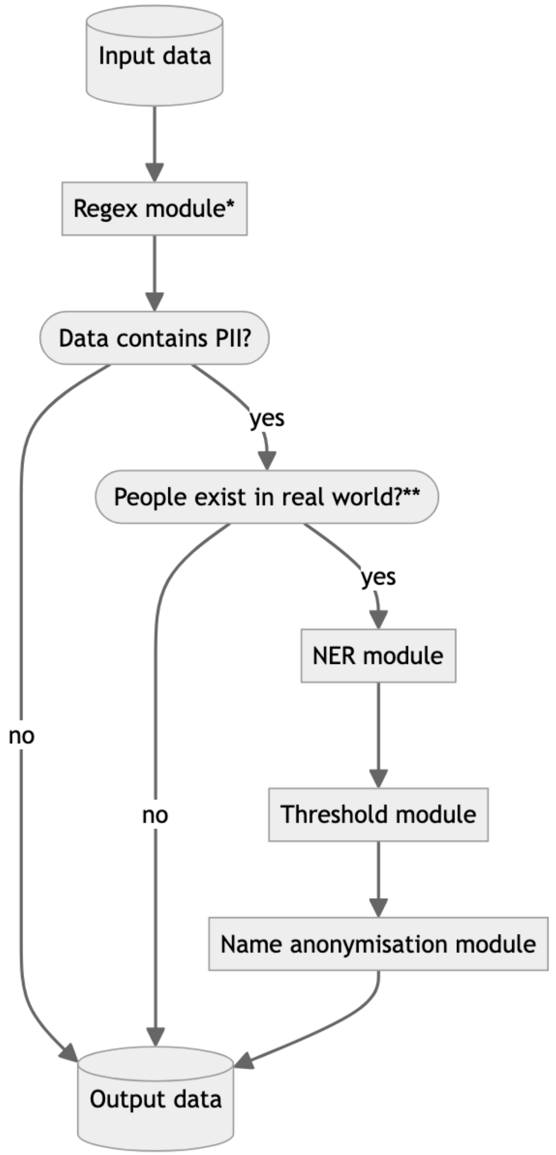
Data curation: PII

Tension between following two values:

1. We want the model to **learn as much as possible** about the world.
   Therefore, we do not want to remove information that is crucial to achieve this (e.g. names of well-known people, well-known street names, etc.)

2. We want to **remove** as much **personally identifiable information** as possible

Solution: very thorough triage



Data curation: bias

Bias in ChatGPT:

Write a haiku about men

Mountains stand with pride,
Steady in the storm's embrace,
Strength in quiet calm.

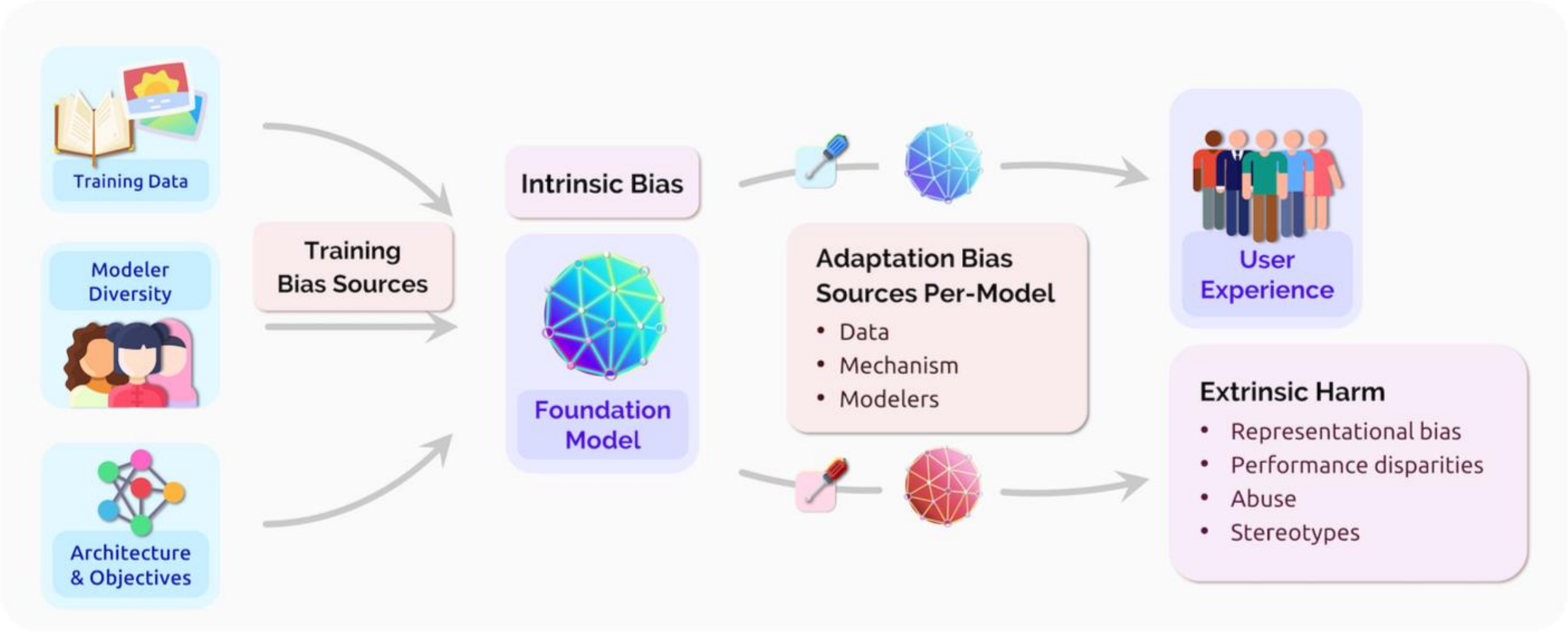
Oaks reach to the sky,
Roots deep in the earth below,
Silent strength they hold.

Write a haiku about women

Moonlight softly glows,
Petals bloom in midnight's hush,
Grace in every step.

Rivers carve their path,
Flowing with both might and grace,
Endless, they endure.

Data curation: bias



Source: 'On the Opportunities and Risks of Foundation Models', Center for Research on Foundation Models (CRFM), Stanford Institute for Human-Centered Artificial Intelligence (HAI), Stanford University

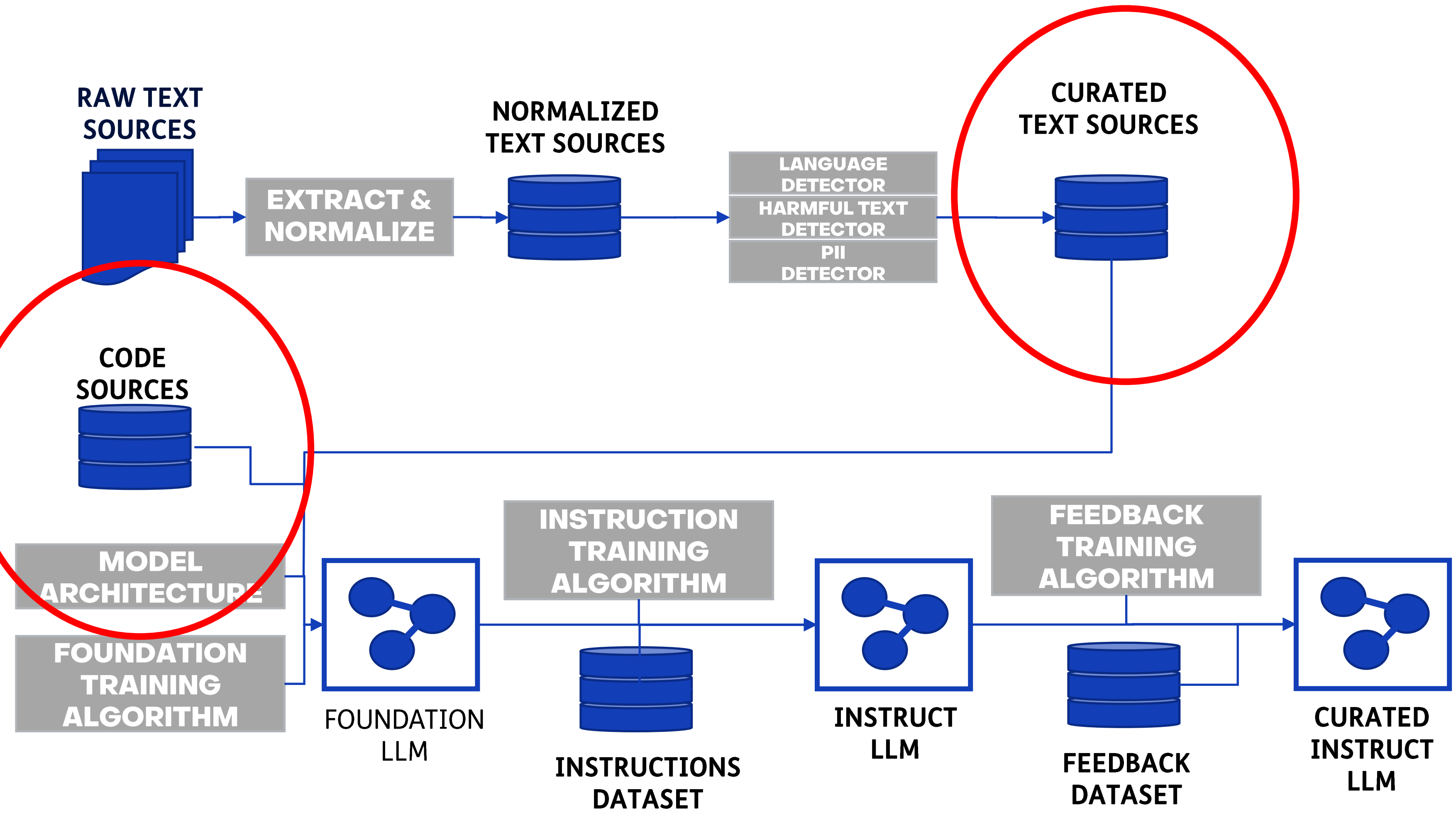
Data curation: bias

Representation bias: Is there enough diversity in the data?

- Determining which groups to include (e.g. people of different genders, ethnicities, religions)
- Applying a method to identify representation in the input data for these groups
- Using this information to:
  1. Be transparent about any representation bias in the data
  2. Try to gather more data on underrepresented groups

GPT-NL

The road to GPT-NL



Dataset

- 300B tokens
- Natural language: 50% Dutch, 50% English
- 60% natural language, 40% high quality code [1]
  - Enhances reasoning

- High quality data outweighs more data [2][3][4]
  - Informative, clear, self-contained, instructive [3]

- [1] Ma et al, At Which Training Stage Does Code Data Help LLM Reasoning? (2024)
- [2] Tan & Wang, 1.5-Pints Technical Report: Pretraining in Days, Not Months (2024)
- [3] Gunasekar et al., Textbooks Are All You Need (2023)
- [4] Sachdeva et al., How to Train Data-Efficient LLMs (2024)

Data desert

- Fewer resources than big tech
    - Compute
    - But especially data

Dall-E imagines a data desert

Data desert

- Fewer resources than big tech
  - Compute
  - But especially data

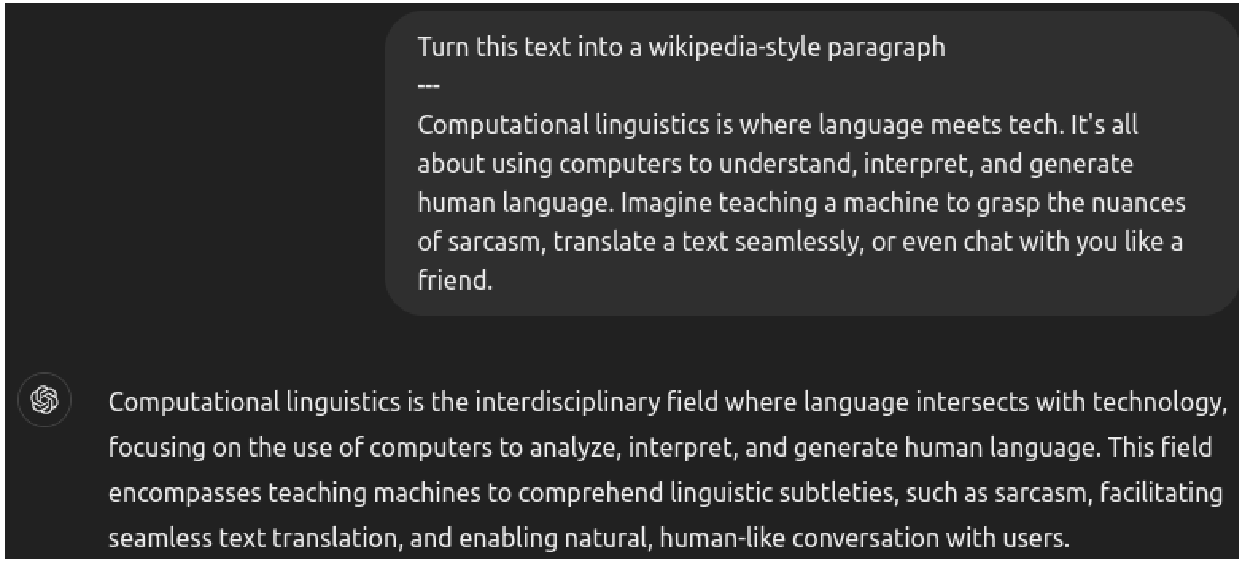
- Overcome with three strategies
  - Synthesis
  - Oversampling
  - Larger model size

Dall-E imagines a data desert

Strategy 1: Data synthesis

- Style transfer
- Machine translation
- Structured data to text data
- Rewriting data
- Based on external data
  - No knowledge extraction from another LLM

- Generated with an LLM that is as compliant as possible



ChatGPT's take on Wikipedia-style text

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

Strategy 2: Oversampling

- Multiple training epochs on the same data
- Limited efficiency loss up to 4 times

- Results diminish
  - Still worth considering up to 30-40 times [1]

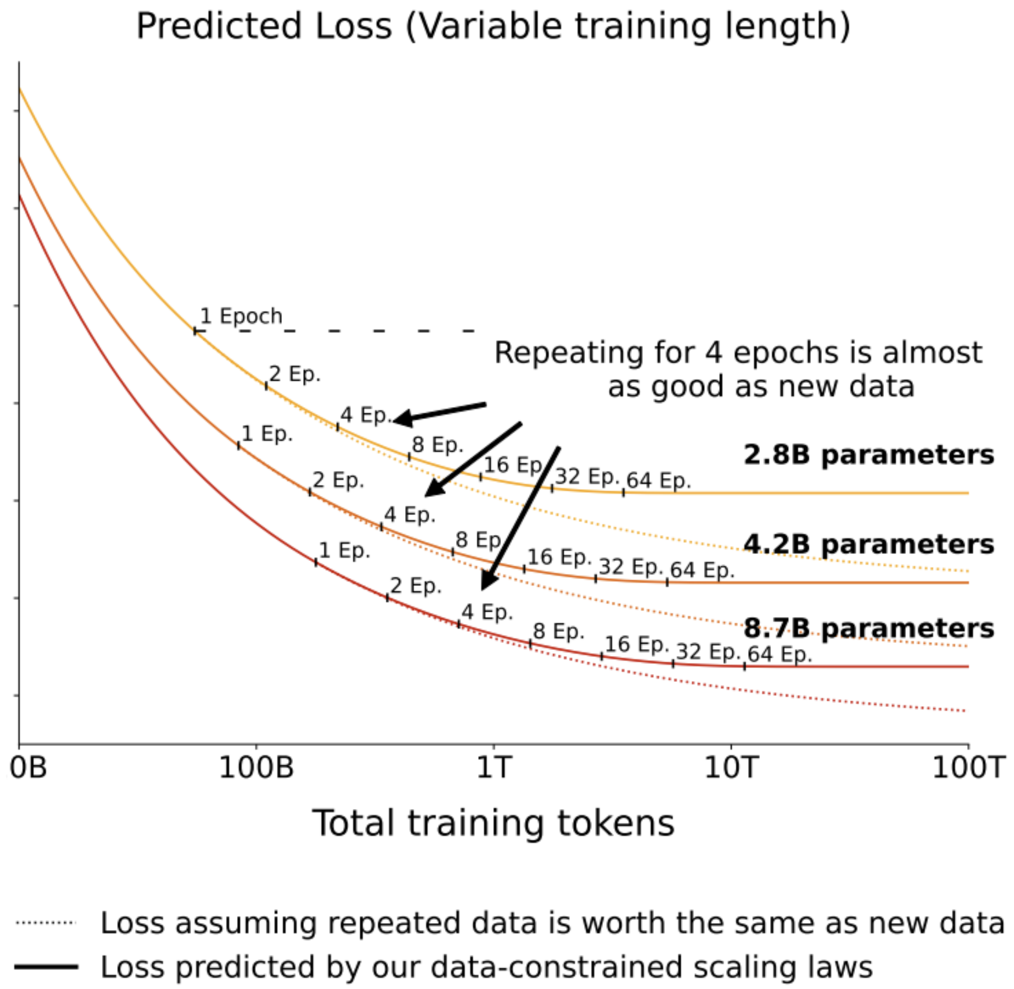
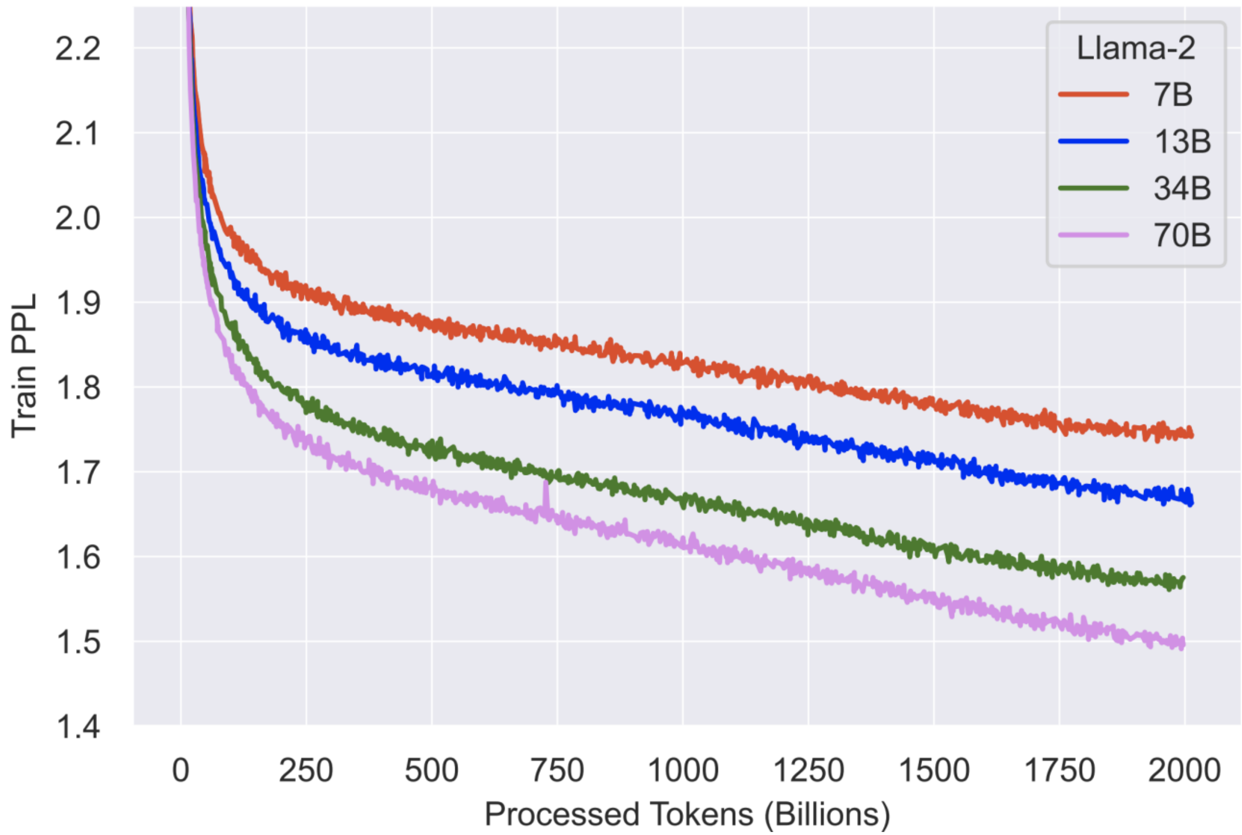


Figure from [1]

[1] Muennighoff et al., Scaling Data-Constrained Language Models (2023)

Strategy 3: Larger model size

- Larger models are smarter
  - With same number of processed tokens
  - But… costlier for inference



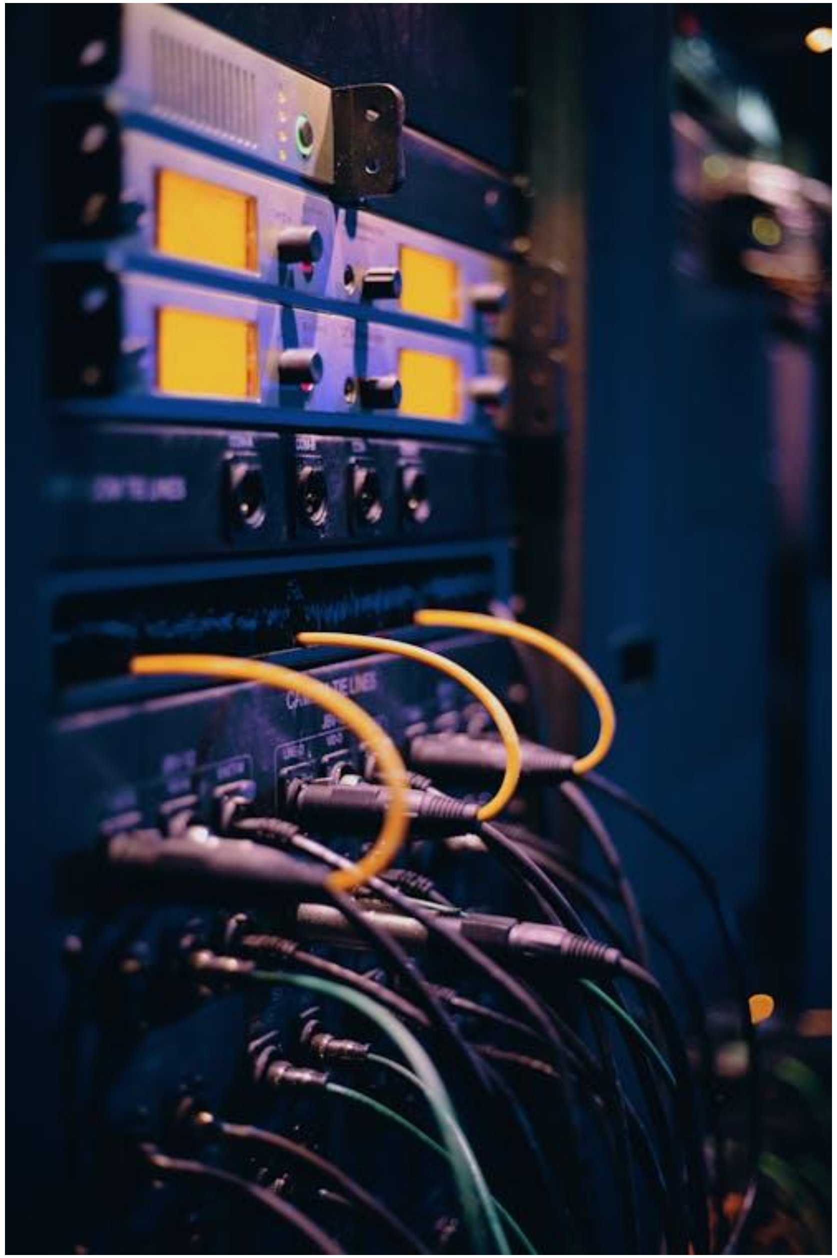
Touvron et al, Llama 2: Open Foundation and Fine-Tuned Chat Models (2023)

Perplexity, lower is better

Model size

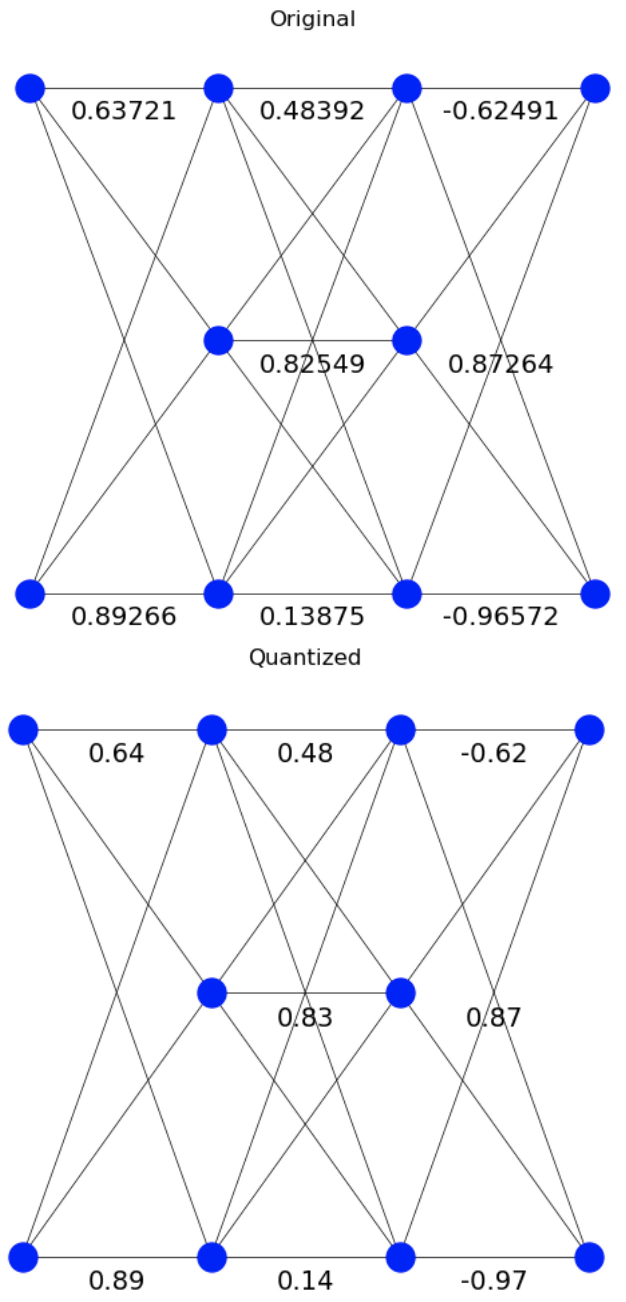
- Model needs to run on professional hardware
  - No focus on consumer hardware
  - But "reasonable" hardware
  - Reasonable: single server-grade GPU

- Keeping in mind energy consumption



Model size

- Training in full precision (BF16)
- Quantizing, when generating, to int-4
  - Only a small performance drop
  - Outperforms models with less parameters
  - At about half of the memory [1]
  - **Train a model twice as large!**

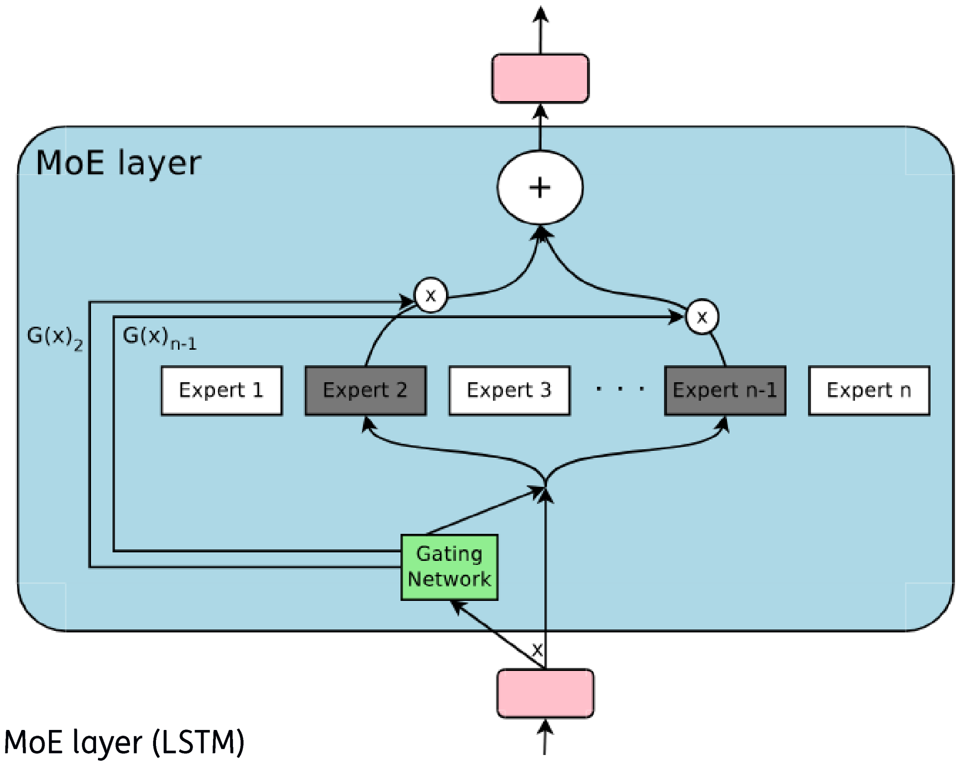


[1] Jin et al., A Comprehensive Evaluation of Quantization Strategies for Large Language Models (2024)

Model size

- Exploring using Mixture of Experts
- Large model will be 8 times the smaller model
- 2 active experts

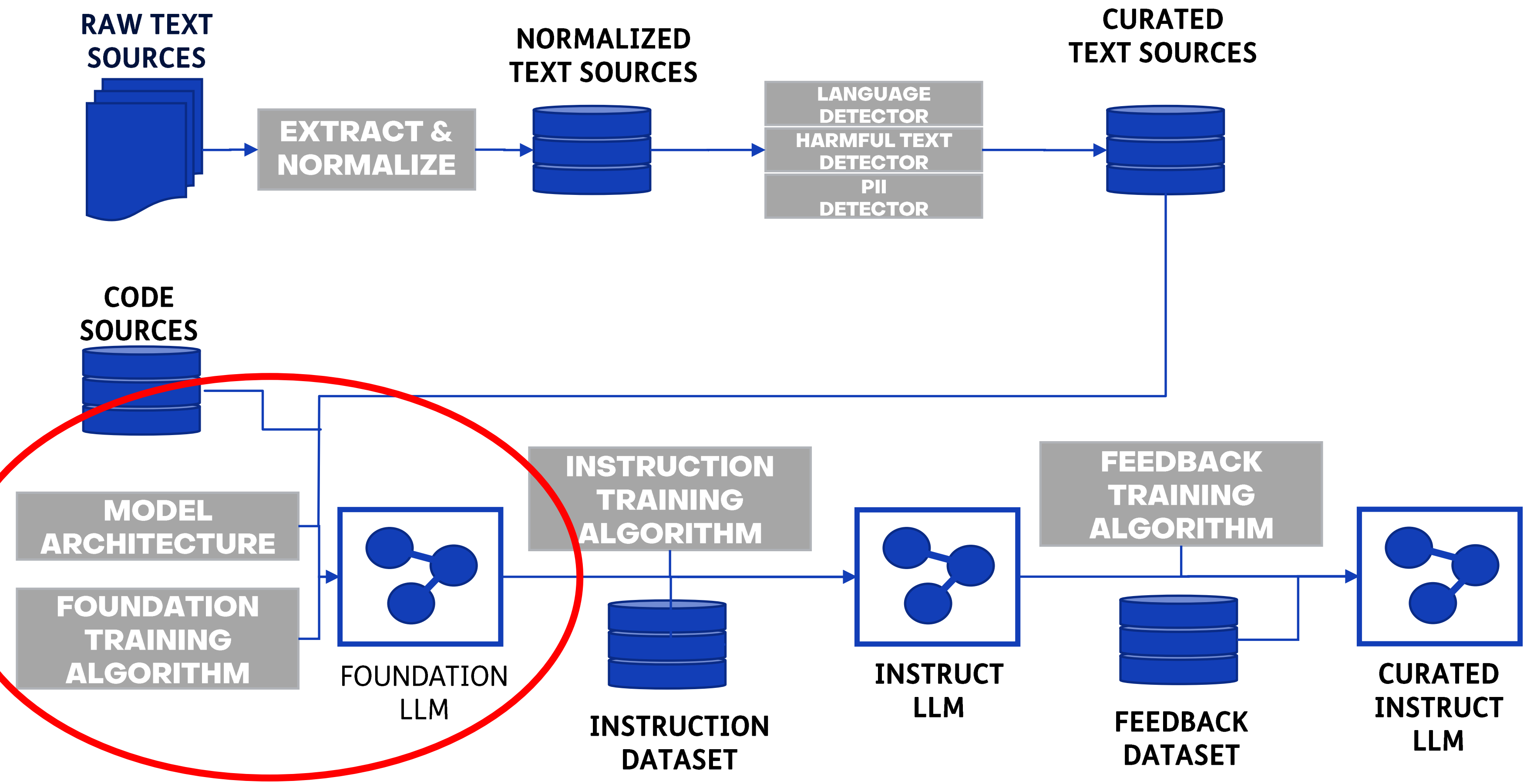
- More performant with a lower parameter count



MoE layer (LSTM)

Shazeer et al., Outrageously Large Neural Networks (2017)

The road to GPT-NL



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

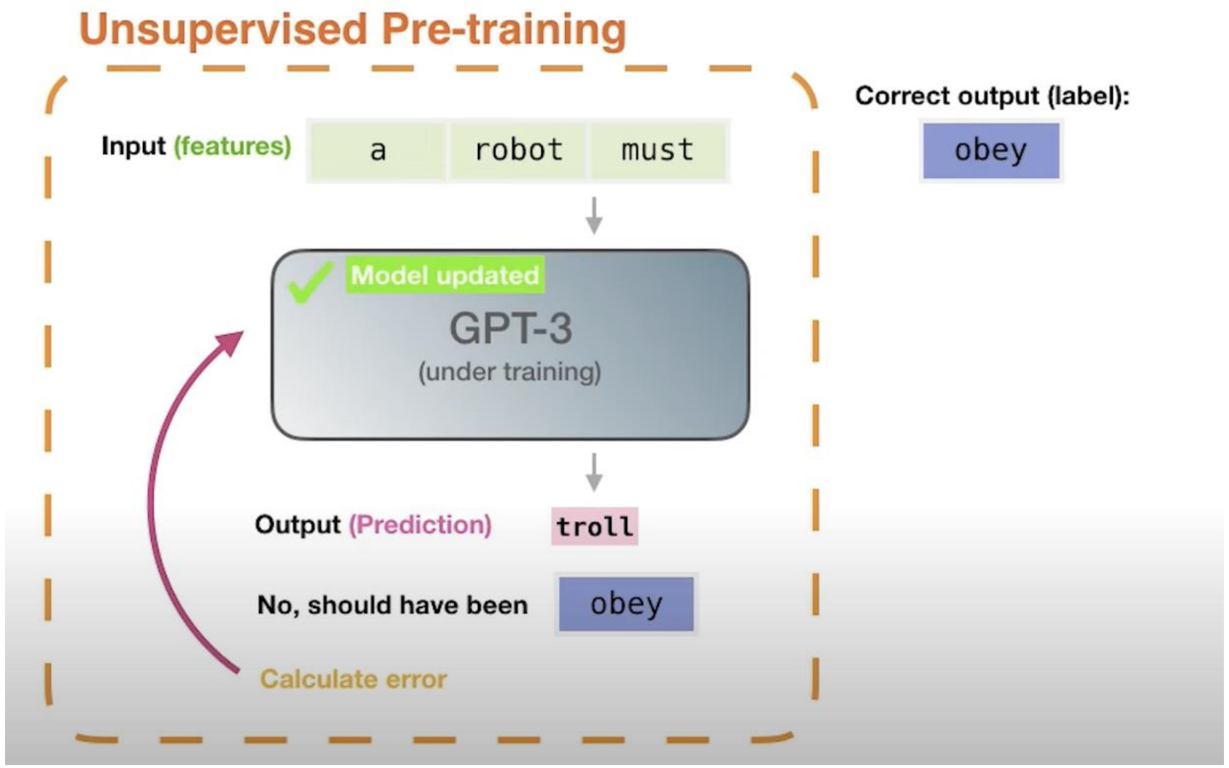
Training hardware

- SURF's Snellius, the Dutch national supercomputer
- GPT-NL has access to 22 H100 96GB GPUs

Training LLM 101

- **Objective**: Minimise the loss of the model towards the training data
  - A low loss means a good understanding of the data distribution

- Given input tokens, predict the next token
- Update the model weights to predict a little bit better next time



Source: https://www.youtube.com/@arp_ai

Architecture

- We are training from scratch
- Basing on Llama (3)'s architecture
  - Openly available
  - Great performance

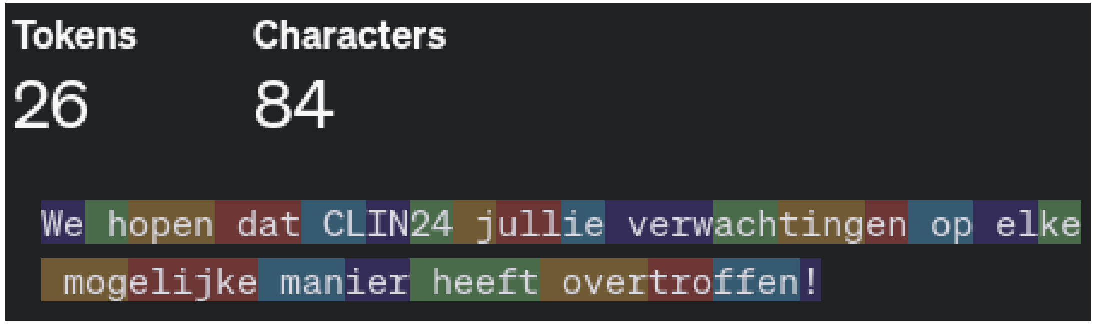
- Final decision to come closer to training
  - Allowing us to adapt to the latest and greatest

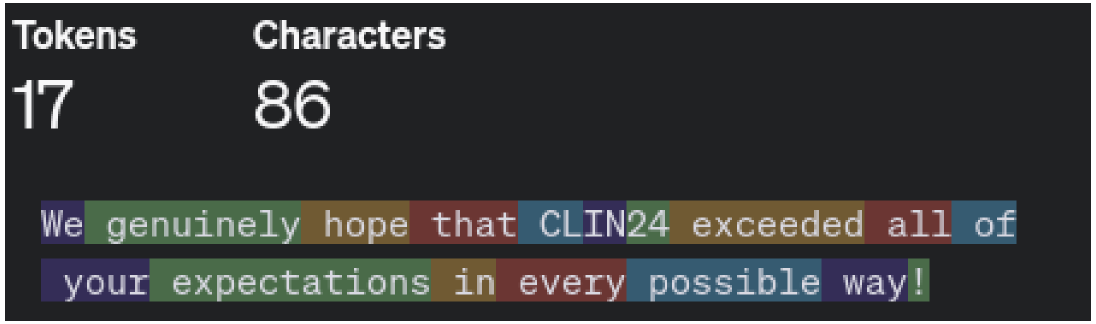
Source: https://github.com/meta-llama/llama3

Tokenizer

- LLMs see tokens rather than letters
- Tokenizers have a vocabulary size (~50k)
- Common tokenizers prioritize English
  - Those tokenizers require more tokens for Dutch
  - More expensive
  - More compute

- We need to train **our own tokenizer**, that fits our dataset





GPT-3.5 & GPT-4 tokenizer sample

https://platform.openai.com/tokenizer

Training frameworks

- Many models are "open source", but training code is rarely available
  - Luckily some implementations are, such as OLMo

PyTorch + FSDP (OLMo-based)

Low-level API
Very customizable

FSDP for distributed training

PyTorch

Transformers + DeepSpeed

High-level API
Based on Transformers

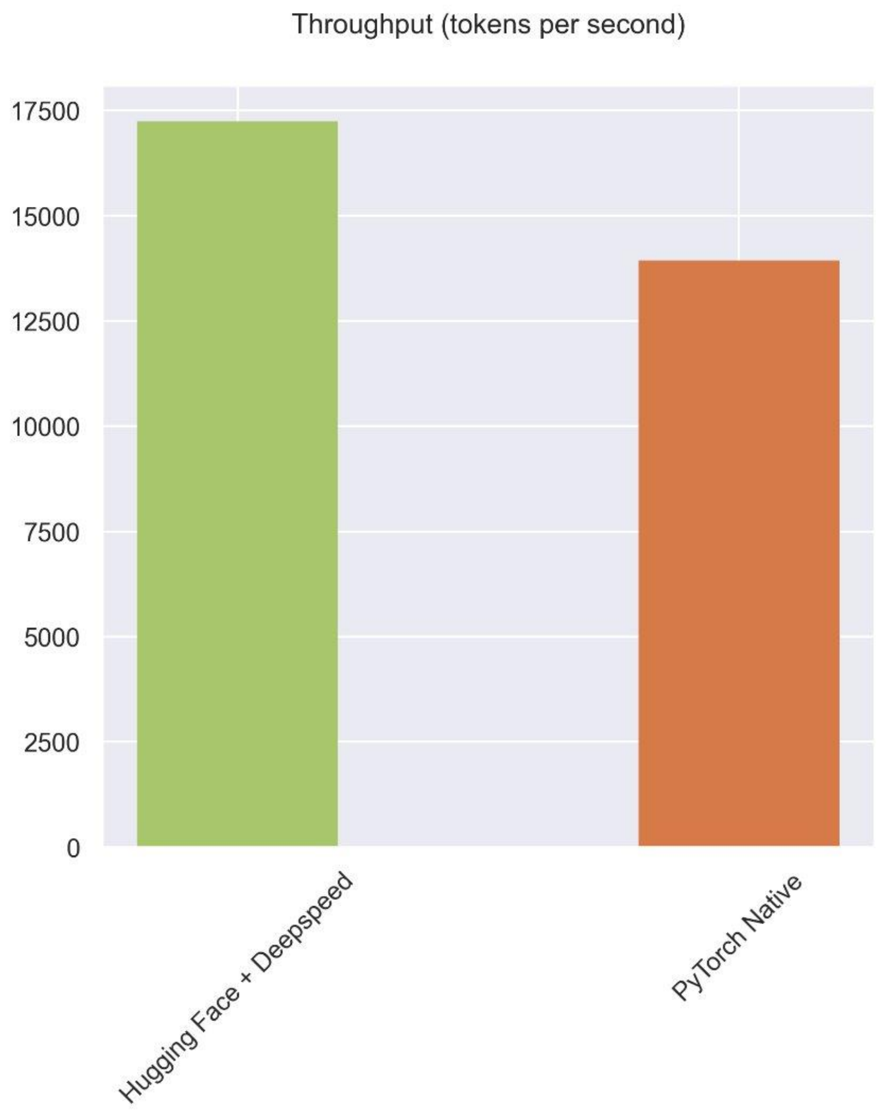
Lots of functionality implemented out of the box
DeepSpeed for distributed training

Hugging Face deepspeed

Training learnings

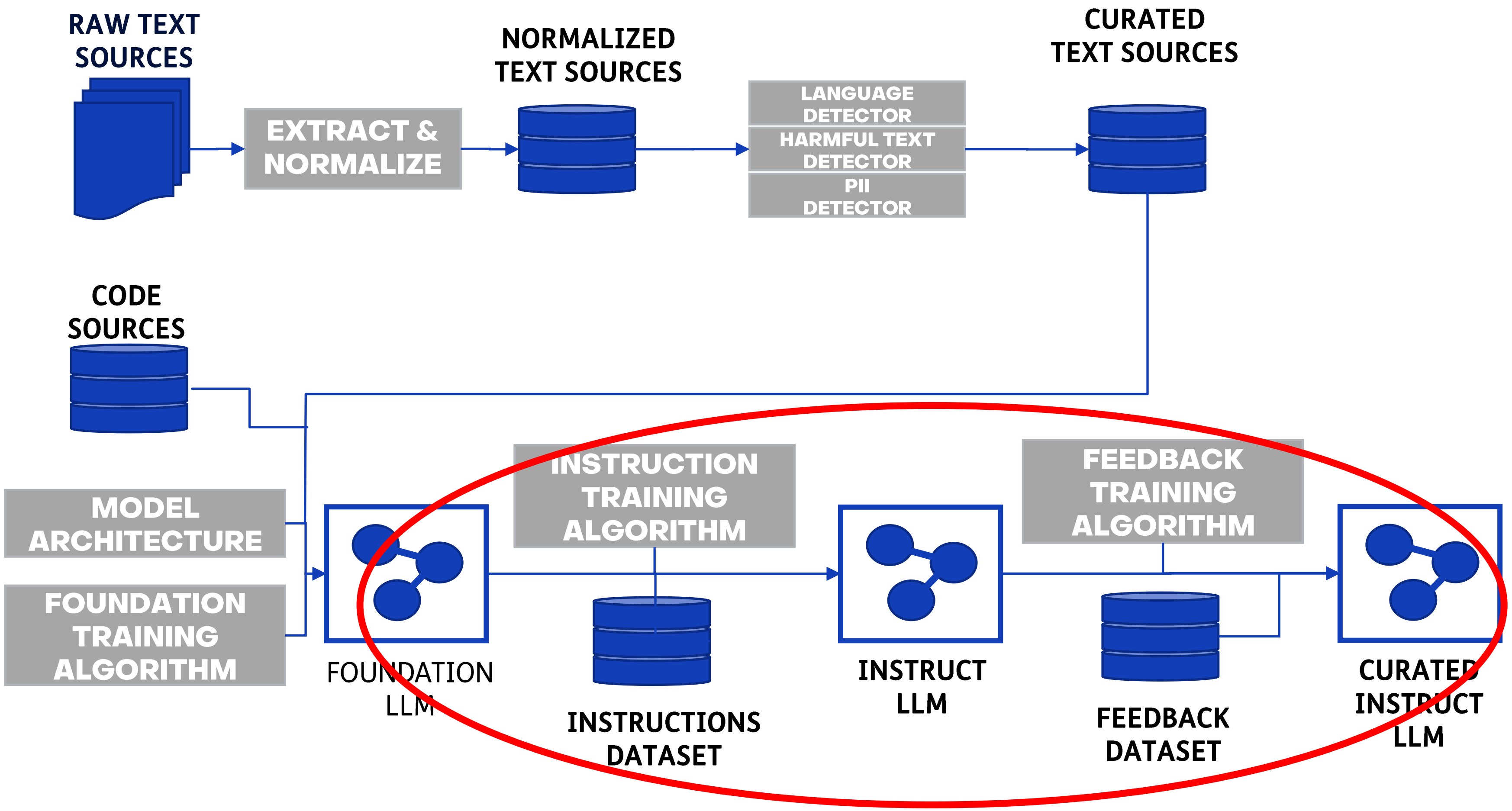
- Without optimization, Deepspeed performs better
- However, still only ±37% of theoretical performance (MFU)
  - On 2 nodes, 8 GPUs
  - Llama-3 reaches ±43% on 8k GPUs

- Further optimization is necessary, especially for PyTorch



GPT-NL

The road to GPT-NL



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

Instruction fine-tuning

- Making the model follow chats and instructions
- High quality English datasets are available
- Dutch is lacking

- Outsourcing creation of Dutch datasets to various annotation companies
  - High quality
  - No machine translation
  - Different kinds of companies
- Starting out with 5k instructions

Type	Dataset Name	# of Instances	# of Tasks	# of Lang	Construction	Open-source
Generalize to unseen tasks	UnifiedQA (Khashabi et al., 2020)[1]	750K	46	En	human-crafted	Yes
	OIG (LAION.ai, 2023)[2]	43M	30	En	human-model-mixed	Yes
	UnifiedSKG (Xie et al., 2022)[3]	0.8M	-	En	human-crafted	Yes
	Natural Instructions (Honovich et al., 2022)[4]	193K	61	En	human-crafted	Yes
	Super-Natural Instructions (?)[5]	5M	76	55 Lang	human-crafted	Yes
	P3 (Sanh et al., 2021)[6]	12M	62	En	human-crafted	Yes
	xP3 (Muennighoff et al., 2022)[7]	81M	53	46 Lang	human-crafted	Yes
	Flan 2021 (Longpre et al., 2023)[8]	4.4M	62	En	human-crafted	Yes
	COIG (Zhang et al., 2023a)[9]	-	-	-	-	Yes
Follow users' instructions in a single turn	InstructGPT (Ouyang et al., 2022)	13K	-	Multi	human-crafted	No
	Unnatural Instructions (Honovich et al., 2022)[10]	240K	-	En	InstructGPT-generated	Yes
	Self-Instruct (Wang et al., 2022c)[11]	52K	-	En	InstructGPT-generated	Yes
	InstructWild (Xue et al., 2023)[12]	104K	429	-	model-generated	Yes
	Evol-Instruct (Xu et al., 2023a)[13]	52K	-	En	ChatGPT-generated	Yes
	Alpaca (Taori et al., 2023)[14]	52K	-	En	InstructGPT-generated	Yes
	LogiCoT (Liu et al., 2023a)[15]	-	2	En	GPT-4-generated	Yes
	Dolly (Conover et al., 2023)[16]	15K	7	En	human-crafted	Yes
	GPT-4-LLM (Peng et al., 2023)[17]	52K	-	En&Zh	GPT-4-generated	Yes
	LIMA (Zhou et al., 2023)[18]	1K	-	En	human-crafted	Yes
Offer assistance like humans across multiple turns	ChatGPT (OpenAI, 2022)	-	-	Multi	human-crafted	No
	Vicuna (Chiang et al., 2023)	70K	-	En	user-shared	No
	Guanaco (JosephusCheung, 2021)[19]	534,530	-	Multi	model-generated	Yes
	OpenAssistant (Köpf et al., 2023)[20]	161,443	-	Multi	human-crafted	Yes
	Baize v1 (?)[21]	111.5K	-	En	ChatGPT-generated	Yes
	UltraChat (Ding et al., 2023a)[22]	675K	-	En&Zh	model-generated	Yes

[1] https://github.com/allenai/unifiedqa
[2] https://github.com/LAION-AI/Open-Instruction-Generalist
[3] https://github.com/hkunlp/unifiedskg
[4] https://github.com/allenai/natural-instructions-v1
[5] https://github.com/allenai/natural-instructions
[6] https://huggingface.co/datasets/bigscience/P3
[7] https://github.com/bigscience-workshop/xmtf
[8] https://github.com/google-research/FLAN
[9] https://github.com/BAAI-Zlab/COIG
[10] https://github.com/orhonovich/unnatural-instructions
[11] https://github.com/yizhongw/self-instruct
[12] https://github.com/XueFuzhao/InstructionWild
[13] https://github.com/nlpxucan/evol-instruct
[14] https://github.com/tatsu-lab/stanford_alpaca
[15] https://github.com/csitfun/LogiCoT
[16] https://huggingface.co/datasets/databricks/databricks-dolly-15k
[17] https://github.com/Instruction-Tuning-with-GPT-4/GPT-4-LLM
[18] https://huggingface.co/datasets/GAIR/lima
[19] https://huggingface.co/datasets/JosephusCheung/GuanacoDataset
[20] https://github.com/LAION-AI/Open-Assistant
[21] https://github.com/project-baize/baize-chatbot
[22] https://github.com/thunlp/UltraChat#data

Table 1: An overview of instruction tuning datasets.

Zhang et al (2023), Instruction Tuning for Large Language Models: a Survey

Feedback tuning

- Further finetuning the model
  - Fitting human preferences
  - Achieving alignment (helpful, honest, not harmful..)

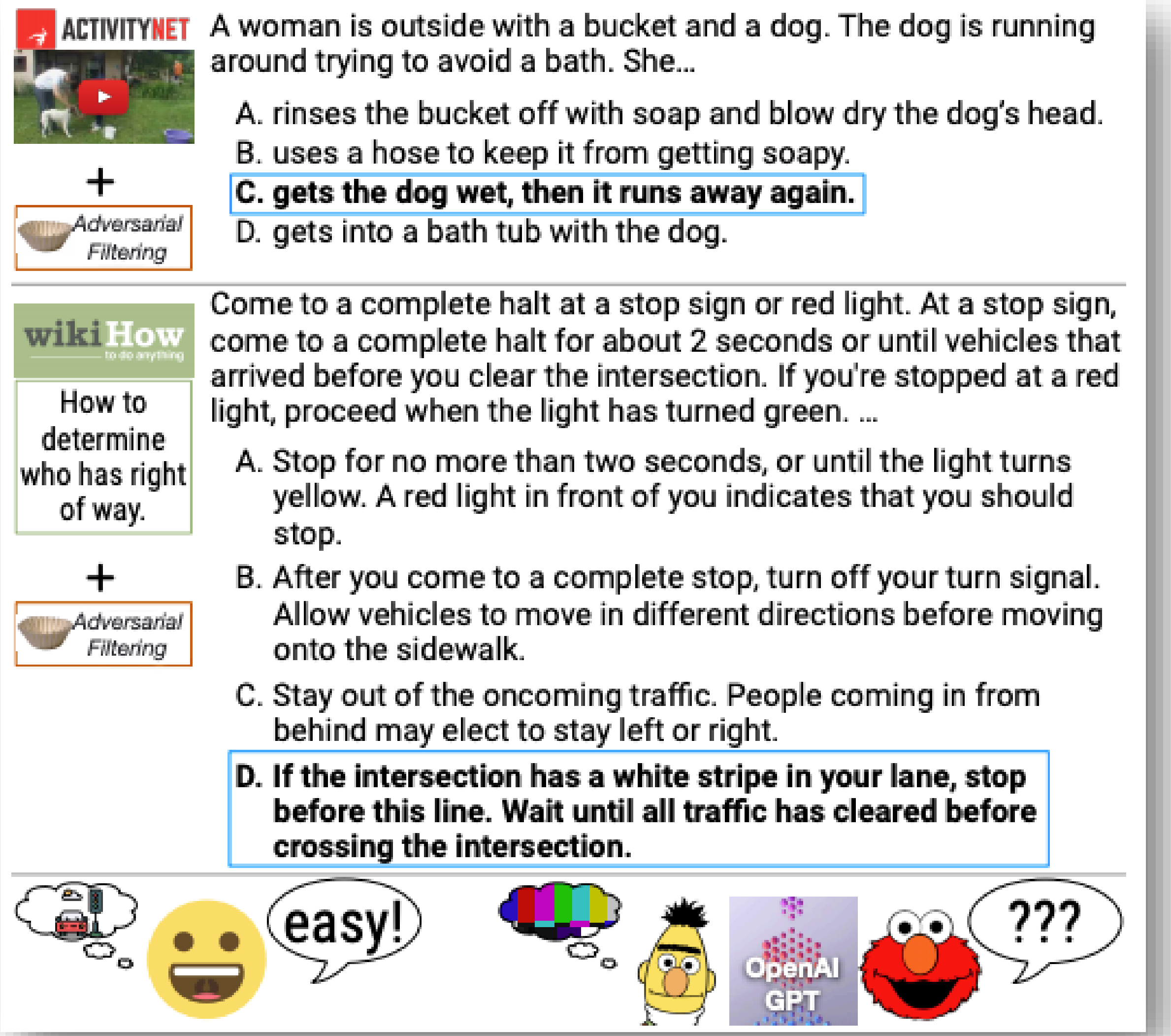
- Focus on aligning to prevent **accidental** harmful content
- No focus on “neutering” the model
  - Reduces performance
  - Those with malicious intentions will prefer other models regardless

A Person Tuning a Bass Guitar by Artem Podrez (Pexels.com)

Evaluation & Benchmarking

Evaluation & Benchmarking

- Dataset-based benchmarking
- Most existing benchmarks are translated
    - Limited Dutch knowledge



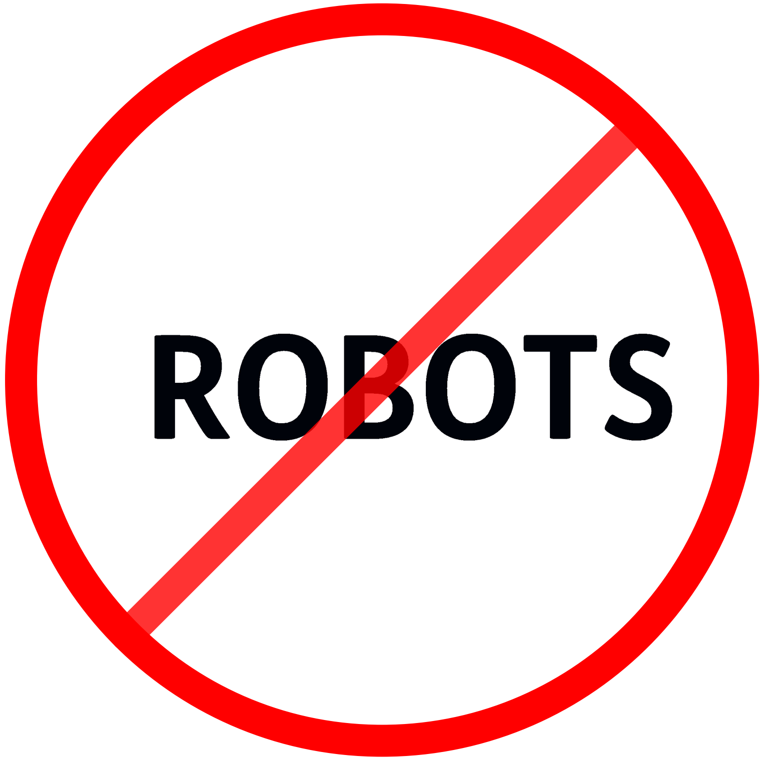
[Hellaswag](https://rowanzellers.com/hellaswag/)
https://rowanzellers.com/hellaswag/

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

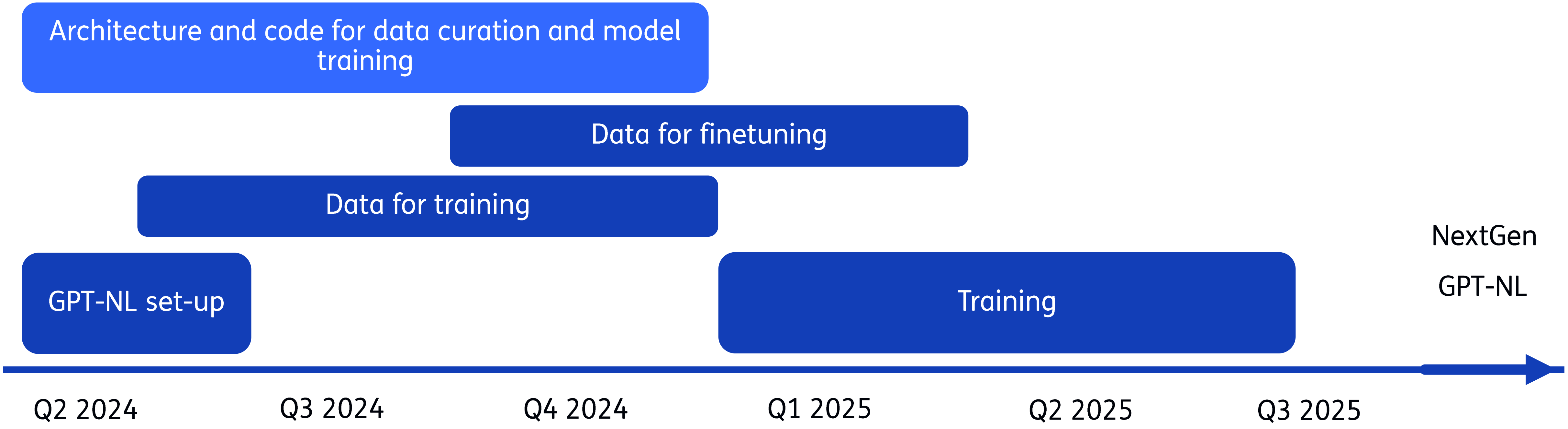
Evaluation & Benchmarking

- Task performance in Dutch
  - Reasoning
  - Instruction following
  - Summarization
  - Simplification
- Dutch cultural understanding and linguistic abilities
  - Bias & inclusion

- No machine translations!



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

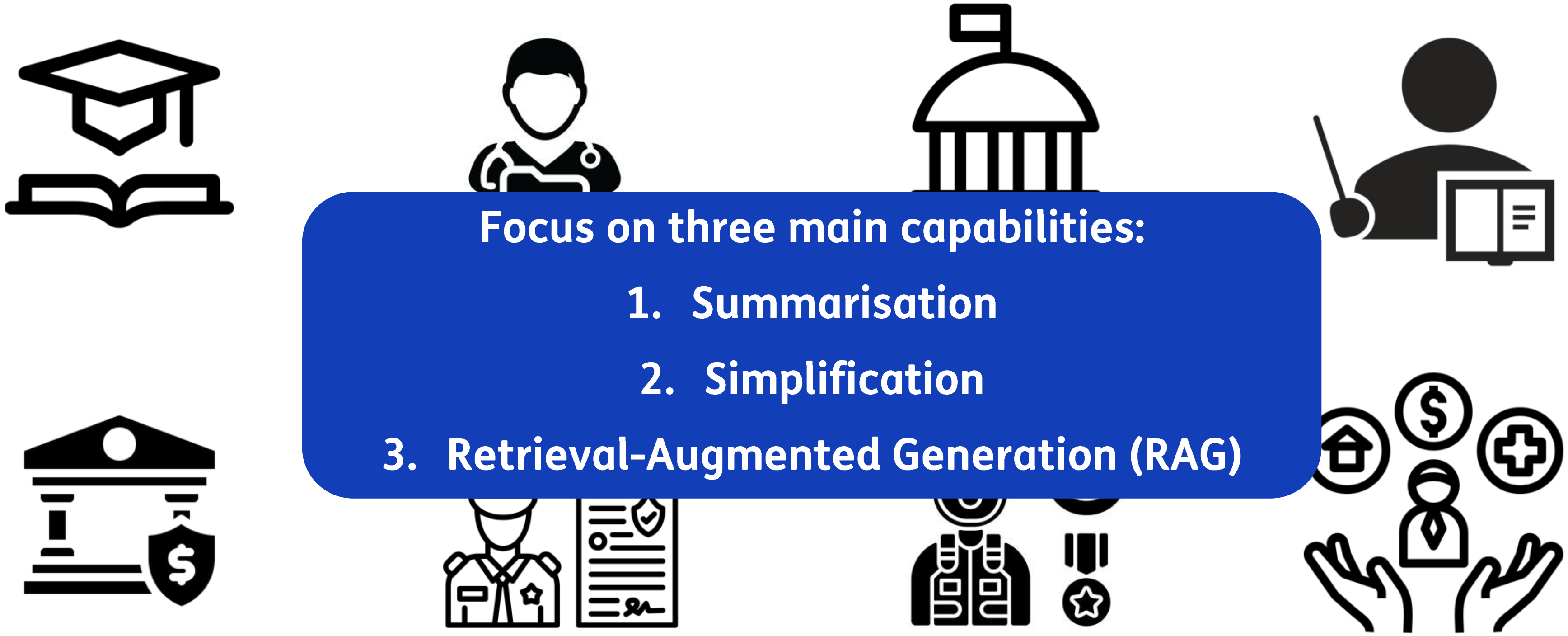
Thank you for your attention!

Dominique Blok dominique.blok@tno.nl

Erik de Graaf erik.degraaf@tno.nl

GPT-NL gpt-nl@tno.nl, gpt-nl.nl

For whom?



Focus on three main capabilities:

1. Summarisation
2. Simplification
3. Retrieval-Augmented Generation (RAG)

Main capabilities and use case

	Summarization	Simplification	RAG
Main capabilities	• Regulations • Compliance requirements	• Simplify complex jargon without compromising on factualness • Language levels specified to user	• Access to and integration of organizational specific (sensitive) information • Provide interface for Q&A to users
Use cases	• Case law documents • Insurance policies • Driving license guidelines • Medicine prescription explanations • Etc.		