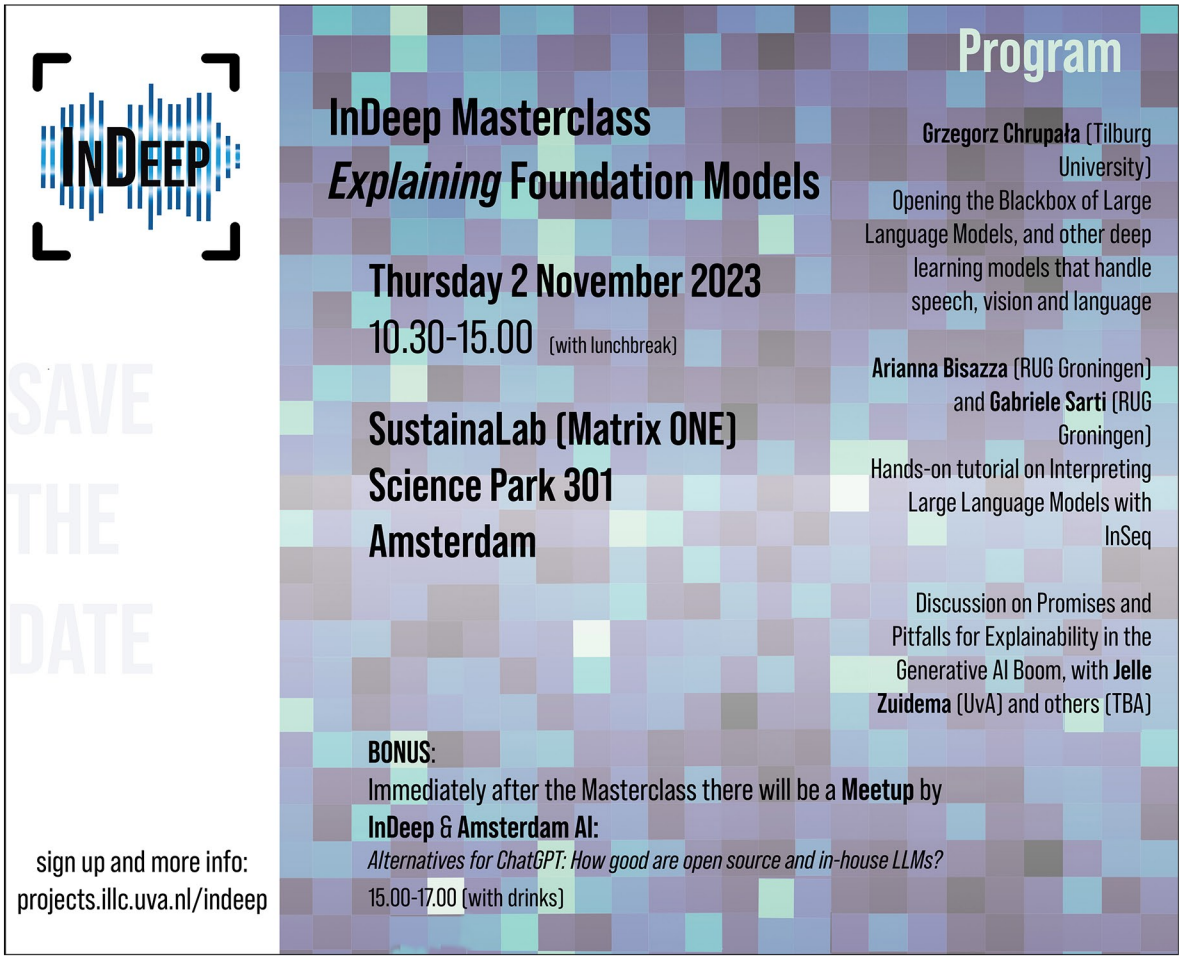
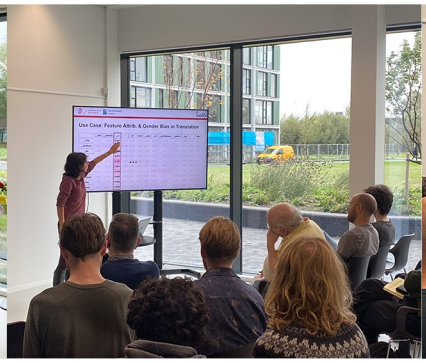


On Thursday 2 November 2023 **InDeep** organized a Masterclass on *Explaining Foundation Models* with presentations, discussion and a hands-on tutorial. After the Masterclass there was a Meetup organized by Indeep and Amsterdam AI


Grzegorz Chrupała (Tilburg University) gave a presentation on opening the Blackbox of Large Language Models and other deep learning models that handle speech, vision and language

Opening the black box of foundation models

Grzegorz Chrupała

What is a foundation model?

- Trained in a self-supervised way (without manual annoations).
- On large amounts of data.
  - Text, audio, images, video
- Starting point for many **downstream** tasks.

How it works

- Language model trained on speech.
- Objective: identify masked audio segment.
- Additionally **tuned** for
  - Transcribe speech into text
  - Recognize spoken commands
  - Identify speakers
  - Classify emotions

Towards **multiple** modalities

- For example audio-visual models learn from images (or videos) accompanied by speech.
- Most recent version of ChatGPT can process text, images, communicate with voice.

Foundation models are the core of AI.

We want AI models to **perform well**.

What other **properties** are **desirable**?

Why do we want to understand models?

- **Explain** decisions to end-users
  - As humans we seek justifications.
  - Arguably, human reasoning is designed for argumentation.
    - We except to AIs to be able to justify themselves the way we do.
  - Justifications of automated decisions may be legally mandated
- **Understand** models as scientists and engineers
  - Knowing when and why models fail gives us insight into
    - The phenomena modeled.
    - Ways to improve model performance

There is a lot of research activity dedicated to **understanding** and **explicating** how **deep learning** works

What feature of DL makes this **interesting** and/or **necessary**?

Distributed representations

- DL learns **distributed** representations.
  - Information is encoded in **patterns** of activation.
  - People are better at grasping information encoded **symbolically**.

Size and complexity

- Foundation models are **large**
  - Millions or billions of learnable parameters
- They are composed of dozens of types of specialized modules
  - Feedforward, Convolutions, Attention, Normalization, Recurrences, ...
- Unlike in many other ML models, the path from input features to outputs is highly **indirect**.

What does it mean to **understand** how a model works?

What kind of **questions** would we like to answer about this?

Which features of an input example led to a particular prediction?

- **Local** explanation, for a particular datapoint.
- Aka **feature attribution**.
- What **minimal change** to this datapoint would cause a change in the prediction?

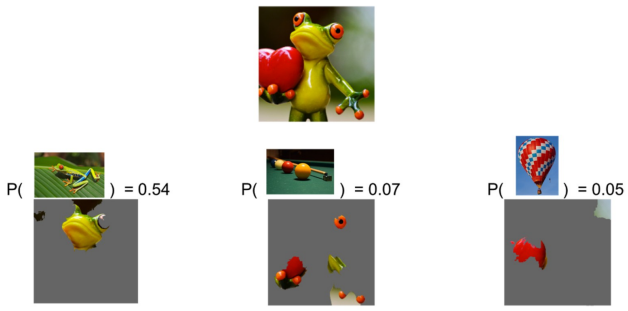


Figure 6. Explanation for a prediction from Inception. The top three predicted classes are "tree frog," "pool table," and "balloon." Sources: Marco Tulio Ribeiro, Pixabay (frog, billiards, hot air balloon).

https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/

Model-agnostic methods

- Occlusion analysis
  - Features or groups of features are occluded and effect on prediction computed.
- LIME – Local Interpretable Model-Agnostic Explanations
- SHAP – Shapley additive explanations

Model-specific methods

- Gradient-based methods
  - Use **gradient** of target function to determine what features affect the output of the model.
- Methods tailored to specific architectures
  - Many recent approaches focus on **Transformers**
  - **ValueZeroing** exploits the mathematical formulation of the transformer to exclude tokens from computation at specific layers (http://dx.doi.org/10.18653/v1/2023.eacl-main.245).

Which input features are important for model decisions?

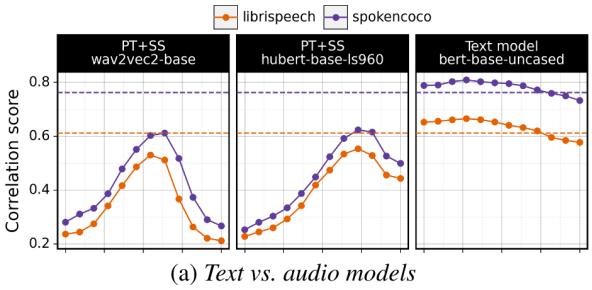
- **Global** explanations, aggregated over multiple datapoints.

Important caveats

- What counts as a feature for the purposes of explanation?
- The use of feature attribution or importance scores assumes a (locally) linear explanatory model.
  - May or may not be a satisfactory approximation to the true function.

What information is encoded in activation patterns?

- What kind of information is encoded?
- To what extent?
- Where is it encoded, in terms of layers, or individual neurons?
- How can we visualize it or otherwise make it comprehensible?



(a) *Text vs. audio models*

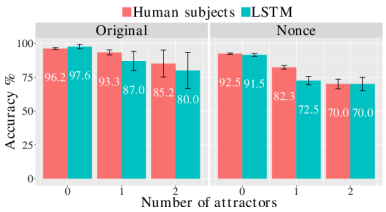
https://dx.doi.org/10.21437/Interspeech.2023-679

How does the model behave when faced with particular phenomena?

- Select or manipulate stimuli to highlight phenomena of interest, and analyze model **behavior** on them.
- Analogous to how a psychologist works.

Subject-verb number agreement

"Yet the ratio of men who survive to the women and children who survive **[is/are]**"



https://doi.org/10.18653/v1%2FN18-1108

Which (symbolic) algorithm does the model implement?

Often analysis is only feasible for a simplified version of an architecture of interest.

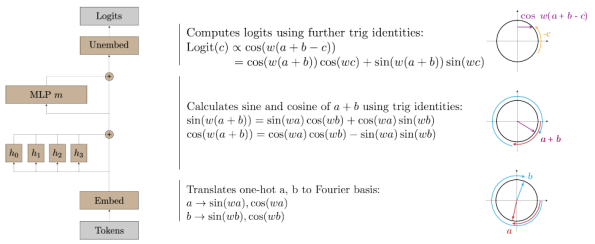


Figure 1: The algorithm implemented by the one-layer transformer for modular addition. Given two numbers a and b, the model projects each point onto a corresponding rotation using its embedding matrix. Using its attention and MLP layers, it then composes the rotations to get a representation of $a+b \mod P$. Finally, it "reads off" the logits for each $c \in \{0, 1, ..., P-1\}$, by rotating by $-c$ to get $\cos(w(a+b-c))$, which is maximized when $a+b \equiv c \mod P$ (since w is a multiple of $\frac{2\pi}{P}$).

https://doi.org/10.48550/arXiv.2301.05217

Evaluation

How can explanations or analyses be **evaluated**?

Main evaluation types

- **Plausibility**
  - does the explanation seem reasonable from the human point of view?
- **Faithfulness**
  - does the explanation accurately capture the factors behind the model's decision?

Explaining foundation models

- Very active area of research.
- Important new developments happen rapidly.
- Many open problems.

Arianna Bisazza (RUG Groningen) and **Gabriele Sarti** (RUG Groningen) gave a hands-on tutorial on interpreting Large Language Models with InSeq

Interpreting (Large) Language Model Generations with Inseq

Gabriele **Sarti** & Arianna **Bisazza**

InDeep Masterclass – November 2nd, 2023

university of groningen / center for language and cognition

Dutch Research Council

university of groningen | Dutch Research Council | INDEEP

About Us

Gabriele Sarti
3nd-year PhD student

Arianna Bisazza
Associate professor

@ University of Groningen

→ InDeep Work Package on improving interpretability for **machine translation** from a user-centric perspective

→ **Core working methods:** Feature attribution for text-generating models

university of groningen | Dutch Research Council | INDEEP

Before we begin, let's set up our notebook:

`https://tinyurl.com/indeep-masterclass-ex`

OR

university of groningen | Dutch Research Council | INDEEP

Feature Attribution

- Understanding why Neural Language Models make certain decisions when generating text is hard!
- A promising approach to do that is Feature Attribution

university of groningen | Dutch Research Council | INDEEP

What is Feature Attribution?

Which part(s) of the input sample contributed most to a model's prediction?

A woman is throwing a frisbee in a park.

A large white bird standing in a forest.

Feature attribution examples for an image captioning model, from [Xu et al. 2016]

university of groningen | Dutch Research Council | INDEEP

What is Feature Attribution?

Which part(s) of the input sample contributed most to a model's prediction?

	$p(y\|\mathbf{x})$	y
the year 's best and most unpredictable comedy	0.91	pos
we never feel anything for these characters	0.95	neg

Sentiment analysis example from [Madsen et al. 2021]

university of groningen | Dutch Research Council | INDEEP

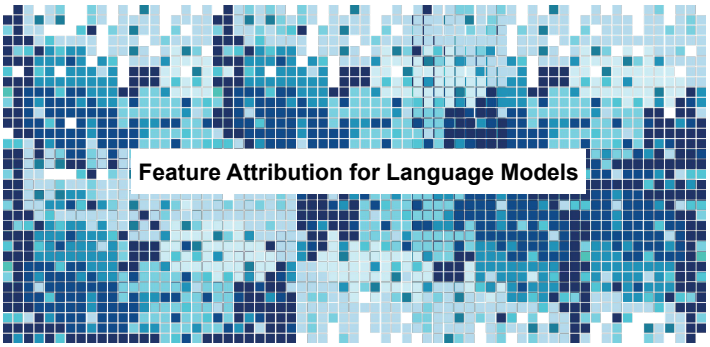
What is Feature Attribution?

Which part(s) of the input sample contributed most to a model's prediction?

	$p(y\|\mathbf{x})$	y
the year 's best and most unpredictable comedy	0.91	pos
we never feel anything for these characters	0.95	neg

Sentiment analysis example from [Madsen et al. 2021]

university of groningen | Dutch Research Council | INDEEP



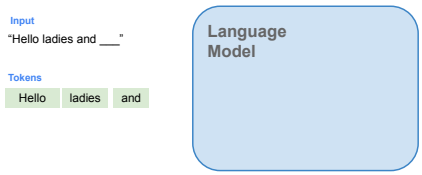
university of groningen | Dutch Research Council | INDEEP

Feature Attribution for Language Models

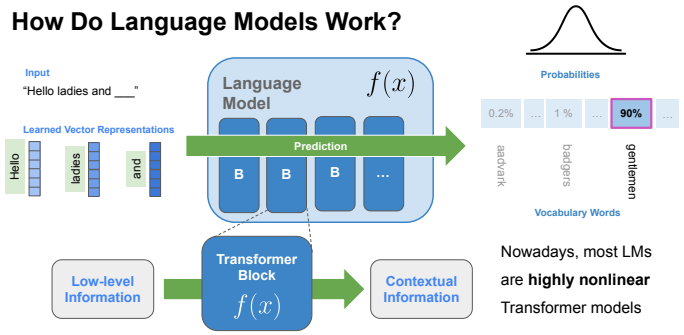
- ⚠️ **Urgent**!
  - More and more NLP tasks (even classification ones) are solved by text-generating models nowadays
  - Understanding the relationship between a model's inputs & output can help us detect unwanted **errors and biases**
- ⚙️ **Challenging**! More difficult than classical classification
  - Large output space
  - At each timestep *t*, previously generated words affect prediction (i.e. input features are *dynamic*)

university of groningen | Dutch Research Council

How Do Language Models Work?

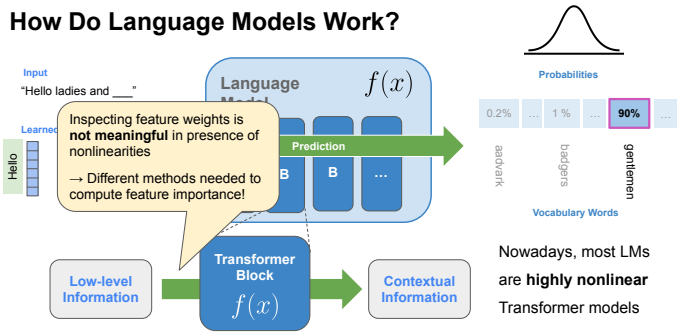


How Do Language Models Work?

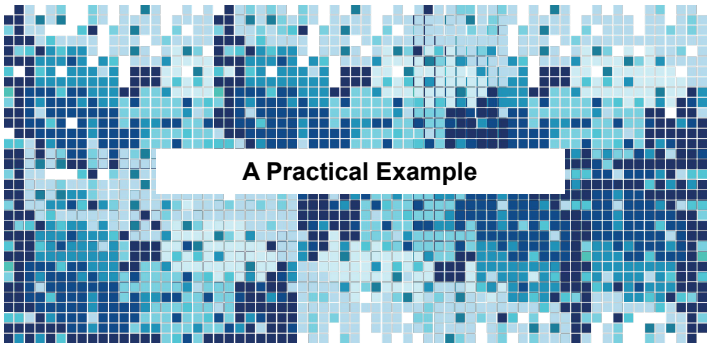


Nowadays, most LMs are **highly nonlinear** Transformer models

How Do Language Models Work?



Nowadays, most LMs are **highly nonlinear** Transformer models



Use Case: Feature Attrib. & Gender Bias in Translation

♂ or ♀ ?

English: The children do not like the teacher.

French: Les enfants n'aiment pas **le professeur**.

♂

Use Case: Feature Attrib. & Gender Bias in Translation

~~♂ or ♀~~

English: The children do not like the teacher because **he** gives a lot of homework.

French: Les enfants n'aiment pas **le professeur** parce qu'**il** donne beaucoup de devoirs.

Use Case: Feature Attrib. & Gender Bias in Translation

~~♂ or ♀~~

English: The children do not like the teacher because **she** gives a lot of homework.

⚠ **Biased prediction**

French: Les enfants n'aiment pas **le professeur** parce qu'**elle** donne beaucoup de devoirs.

Why does that happen??

university of groningen | Dutch Research Council | INDEEP

Use Case: Feature Attrib. & Gender Bias in Translation

	_Les	_enfants	_n	'	ai	ment	_pas	_le	_professeur	_parce	_qu	'	elle
_The	0.047	0.046	0.027	0.02	0.022	0.011	0.016	0.014	0.018	0.019	0.016	0.011	0.013
_children	0.071	0.195	0.057	0.068	0.044	0.021	0.037	0.043	0.05	0.032	0.04	0.025	0.026
_do	0.039	0.036	0.072	0.031	0.059	0.032	0.038	0.01	0.02	0.029	0.021	0.02	0.01
_not	0.077	0.05	0.072	0.056	0.066	0.035	0.078	0.014	0.038	0.041	0.039	0.029	0.015
_like	0.036	0.042	0.161	0.033	0.184	0.074	0.068	0.024	0.02	0.034	0.023	0.027	0.02
_the	0.04	0.042	0.081	0.039	0.069	0.032	0.034	0.091	0.066	0.04	0.028	0.022	0.033
_teacher	0.061	0.103	0.104	0.05	0.076	0.033	0.048	0.314	0.256	0.05	0.042	0.034	0.051
_because	0.04	0.026	0.052	0.035	0.032	0.017	0.031	0.038	0.028	0.164	0.07	0.038	0.057
_she	0.056	0.053	0.033	0.053	0.025	0.016	0.027	0.114	0.052	0.044	0.132	0.044	0.181
_gives	0.04	0.028	0.028	0.029	0.027	0.016	0.02	0.015	0.019	0.052	0.081	0.04	0.081

university of groningen | Dutch Research Council | INDEEP

Use Case: Feature Attrib. & Gender Bias in Translation

	_Les	_enfants	_n	'	ai	ment	_pas	_le	_professeur	_parce	_qu	'	elle
_The	0.047	0.046	0.027	0.02	0.022	0.011	0.016	0.014	0.018	0.019	0.016	0.011	0.013
_children	0.071	0.195	0.057	0.068	0.044	0.021	0.037	0.043	0.05	0.032	0.04	0.025	0.026
_do	0.039	0.036	0.072	0.031	0.059	0.032	0.038	0.01	0.02	0.029	0.021	0.02	0.01
_not	0.077	0.05	0.072	0.056	0.066	0.035	0.078	0.014	0.038	0.041	0.039	0.029	0.015
_like	0.036	0.042	0.161	0.033	0.184	0.074	0.068	0.024	0.02	0.034	0.023	0.027	0.02
_the	0.04	0.042	0.081	0.039	0.069	0.032	0.034	0.091	0.066	0.04	0.028	0.022	0.033
_teacher	0.061	0.103	0.104	0.05	0.076	0.033	0.048	0.314	0.256	0.05	0.042	0.034	0.051
_because	0.04	0.026	0.052	0.035	0.032	0.017	0.031	0.038	0.028	0.164	0.07	0.038	0.057
_she	0.056	0.053	0.033	0.053	0.025	0.016	0.027	0.114	0.052	0.044	0.132	0.044	0.181
_gives	0.04	0.028	0.028	0.029	0.027	0.016	0.02	0.015	0.019	0.052	0.081	0.04	0.081

university of groningen | Dutch Research Council | INDEEP

Use Case: Feature Attrib. & Gender Bias in Translation

	_Les	_enfants	_n	'	ai	ment	_pas	_le	_professeur	_parce	_qu	'	elle
_The	0.047	0.046	0.027	0.02	0.022	0.011	0.016	0.014	0.018	0.019	0.016	0.011	0.013
_children	0.071	0.195	0.057	0.068	0.044	0.021	0.037	0.043	0.05	0.032	0.04	0.025	0.026
_do	0.039	0.036	0.072	0.031	0.059	0.032	0.038	0.01	0.02	0.029	0.021	0.02	0.01
_not	0.077	0.05	0.072	0.056	0.066	0.035	0.078	0.014	0.038	0.041	0.039	0.029	0.015
_like	0.036	0.042	0.161	0.033	...	0.074	0.068	0.024	0.02	0.034	0.023	0.027	0.02
_the	0.04	0.042	0.081	0.039	0.069	0.032	0.034	0.091	0.066	0.04	0.028	0.022	0.033
_teacher	0.061	0.103	0.104	0.05	0.076	0.033	0.048	0.314	0.256	0.05	0.042	0.034	0.051
_because	0.04	0.026	0.052	0.035	0.032	0.017	0.031	0.038	0.028	0.164	0.07	0.038	0.057
_she	0.056	0.053	0.033	0.053	0.025	0.016	0.027	0.114	0.052	0.044	0.132	0.044	0.181
_gives	0.04	0.028	0.028	0.029	0.027	0.016	0.02	0.015	0.019	0.052	0.081	0.04	0.081

university of groningen | Dutch Research Council | INDEEP

Use Case: Feature Attrib. & Gender Bias in Translation

	_Les	_enfants	_n	'	ai	ment	_pas	_le	_professeur	_parce	_qu	'	elle
_The	0.047	0.046	0.027	0.02	0.022	0.011	0.016	0.014	0.018	0.019	0.016	0.011	0.013
_children	0.071	0.195	0.057	0.068	0.044	0.021	0.037	0.043	0.05	0.032	0.04	0.025	0.026
_do	0.039	0.036	0.072	0.031	0.059	0.032	0.038	0.01	0.02	0.029	0.021	0.02	0.01
_not	0.077	0.05	0.072	0.056	0.066	0.035	0.078	0.014	0.038	0.041	0.039	0.029	0.015
_like	0.036	0.042	0.161	0.033	0.184	0.074	0.068	0.024	0.02	0.034	0.023	0.027	0.02
_the	0.04	0.042	0.081	0.039	0.069	0.032	0.034	0.091	0.066	0.04	0.028	0.022	0.033
_teacher	0.061	0.103	0.104	0.05	0.076	0.033	0.048	0.314	0.256	0.05	0.042	0.034	0.051
_because	0.04	0.026	0.052	0.035	0.032	0.017	0.031	0.038	0.028	0.164	0.07	0.038	0.057
_she	0.056	0.053	0.033	0.053	0.025	0.016	0.027	0.114	0.052	0.044	0.132	0.044	0.181
_gives	0.04	0.028	0.028	0.029	0.027	0.016	0.02	0.015	0.019	0.052	0.081	0.04	0.081

university of groningen | Dutch Research Council | INDEEP

Use Case: Feature Attrib. & Gender Bias in Translation

♂ or ♀ ?

English: The children do not like the pretty teacher because **she** gives a lot of homework.

French: Les enfants n'aiment pas **la jolie enseignante** parce qu'**elle** donne beaucoup de devoirs

university of groningen | Dutch Research Council | INDEEP

Use Case: Feature Attrib. & Gender Bias in Translation

	_Les	_enfants	_n	'	ai	ment	_pas	_la	_jolie	_enseignant	e
_The	0.038	0.044	0.033	0.017	0.022	0.01	0.015	0.012	0.009	0.013	0.008
_children	0.062	0.195	0.06	0.061	0.045	0.019	0.036	0.023	0.021	0.046	0.021
_do	0.043	0.039	0.074	0.044	0.058	0.029	0.039	0.014	0.012	0.017	0.01
_not	0.094	0.061	0.085	0.088	0.062	0.034	0.084	0.025	0.023	0.035	0.019
_like	0.032	0.032	0.167	0.038	0.177	0.069	0.068	0.025	0.021	0.02	0.014
_the	0.042	0.038	0.066	0.033	0.062	0.028	0.031	0.077	0.081	0.046	0.024
_pretty	0.06	0.064	0.073	0.059	0.068	0.032	0.054	0.186	0.361	0.074	0.047
_teacher	0.047	0.083	0.066	0.037	0.055	0.026	0.036	0.19	0.119	0.241	0.082
_because	0.036	0.025	0.04	0.038	0.028	0.014	0.026	0.041	0.03	0.028	0.028
_she	0.061	0.05	0.024	0.043	0.021	0.017	0.03	0.108	0.029	0.049	0.038
_gives	0.039	0.027	0.024	0.027	0.024	0.015	0.019	0.027	0.015	0.02	0.016

university of groningen | Dutch Research Council | INDEEP

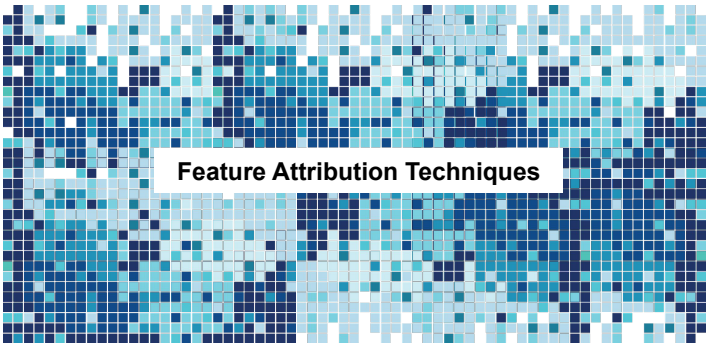
Use Case: Feature Attrib. & Gender Bias in Translation

	p=0.37		p = 0.53
	__le		__le
__The	0.014	__The	0.015
__children	0.043	__children	0.046
__do	0.01	__do	0.01
__not	0.014	__not	0.017
__like	0.024	__like	0.026
__the	0.091	__the	0.093
__teacher	0.314	__teacher	0.349
__because	0.038	__because	0.042
__she	0.114	__he	0.052
__gives	0.015	__gives	0.017

This model seems to pay little attention to the pronoun cue

→ **Unsafe!**

university of groningen | Dutch Research Council | INDEEP



university of groningen | Dutch Research Council | INDEEP

Feature Attribution & Broader Interpretability

- A form of **post-hoc** interpretability (vs. models interpretable by design)
- **Local**: explains a single observation (vs. general model functioning)
- **Low abstraction**: explains model decision in terms of the input features (vs. in terms of relevant training examples; vs. explanations in natural language)

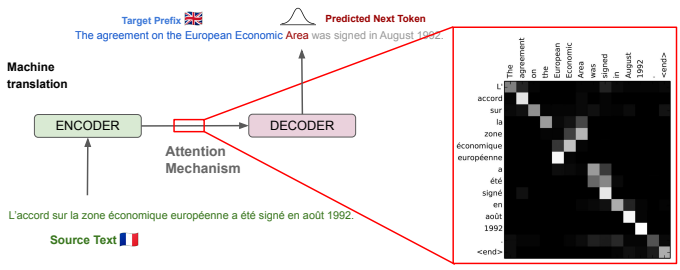
university of groningen | Dutch Research Council | INDEEP

Types of Feature Attribution Methods

- Model-specific (or internals-based)
- Perturbation-based
- Gradient-based

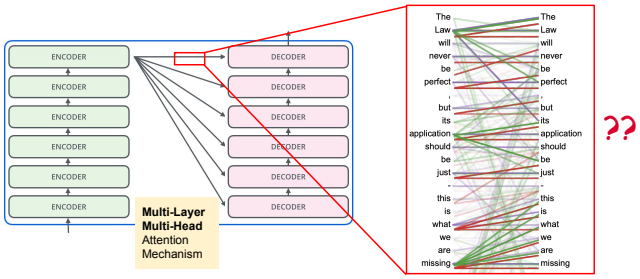
university of groningen | Dutch Research Council | INDEEP

Model Specific: Attention as Attribution



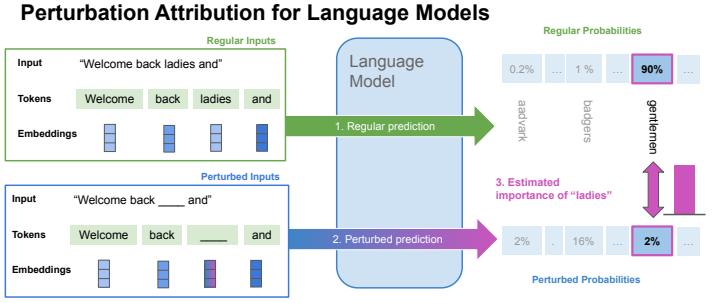
university of groningen | Dutch Research Council | INDEEP

Model Specific: Attention as Attribution in Transformers



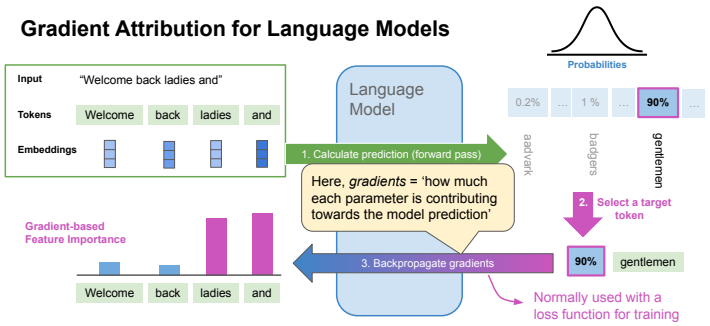
university of groningen | Dutch Research Council | INDEEP

Perturbation Attribution for Language Models



university of groningen | Dutch Research Council | INDEEP

Gradient Attribution for Language Models



university of groningen | Dutch Research Council | INDEEP

In sum: Feature Attribution Methods

- Perturbation-based
- Gradient-based
- Model-specific (or internals-based)

No clear winner (yet)!

All existing attribution methods have some **intrinsic** and/or **practical limitations**

Some methods require **full access** to model internals, other don't

Research is **ongoing**

university of groningen | Dutch Research Council | INDEEP

Inseq : Intepretability for Sequence Generation Models

Pytorch-based Python library to simplify usage and evaluation of **feature attribution** methods for **generative language models**.

Demo @ 61 ACL 2023

215 24

university of groningen

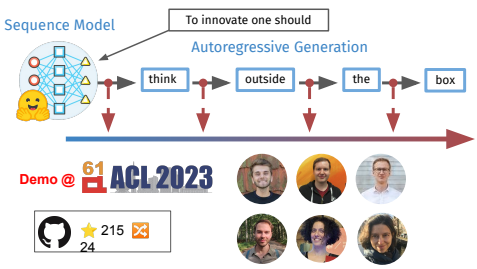
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH | UNIVERSITY OF AMSTERDAM

university of groningen | Dutch Research Council | INDEEP

Inseq: Core Functionalities

university of groningen | Dutch Research Council | INDEEP

Inseq : Interpretability for Sequence Generation Models



Demo @ 61 ACL 2023

215 24

Generated Tokens →

	think	outside	the	box
To	0	0.02	0	0
innovate	0.7	0.45	0.2	0.15
one	0.05	0.04	0	0.02
should	0.25	0.02	0	0.02
think		0.47	0.4	0.15
outside			0.4	0.35
the				0.35
p_{model}	0.5	0.82	0.94	0.99

university of groningen | Dutch Research Council | INDEEP

A Simple Example

```python
import inseq

# Load HF Hub model and attribution method
model = inseq.load_model(
    "google/flan-t5-base",
    "integrated_gradients"
)
# Answer and attribute generation steps
attr_out = model.attribute(
    "Does 3 + 3 equal 6?",
    attribute_target=True
)
# Visualize the generated attribution,
# applying default token-level aggregation
attr_out.show()
```

Source Saliency

	_yes	</s>
_Does	0.264	0.153
_3	0.113	0.099
_+	0.096	0.086
_3	0.096	0.076
_equal	0.213	0.154
_6	0.11	0.108
?	0.106	0.114

Prefix Saliency

	_yes	</s>
_yes		0.21
</s>		

👩‍💻: Does 3 + 3 equal 7? 🤖: yes

Sarti et al. 2023

university of groningen | Dutch Research Council | INDEEP

Main Features

	Method	$f(l)$
G	(Input ×) Gradient	✓
	DeepLIFT	✓
	GradientSHAP	✗
	Integrated Gradients	✓
	Discretized IG	✗
I	Attention Weights	✓
P	Occlusion (Blank-out)	✗
	LIME	✗

G: Gradient-based,
I: Internals-based,
P: Perturbation-based,
f(l): layer-wise attribution

```
>>> model.attribute(
...     "Hello ladies and",
...     generation_args={"max_new_tokens": 9},
...     n_steps=500,
...     internal_batch_size=50
... ).show()
```

- 🤗 Seq2SeqLM and CausalLM support
- Aggregate, save and load attributions
- Batch-attribute datasets in Jupyter & CLI

Sarti et al. 2023

university of groningen | Dutch Research Council | INDEEP

Contrastive Gradient Attribution

Input: *Can you stop the dog from*
Output: barking

1. Why did the model predict "barking"?
Can you stop the dog from

2. Why did the model predict "barking" *instead of* "crying"?
Can you stop the dog from

Disentangle **fluency** factors from the overall model decision, more aligned with human intuition (⬆️ simulatability)

Gradient Attribution

$$\nabla_{y<t}\, q(y_t|y_{<t})$$

e.g.**barking**

Contrastive Gradient Attribution

$$\nabla_{y<t}\left(q(y_t|y_{<t}) - q(\tilde{y}_t|y_{<t})\right)$$

e.g.**barking** e.g.**crying**

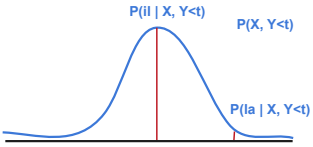
Yin and Neubig, 2022

university of groningen | Dutch Research Council | INDEEP

Step Functions and Custom Attribution Targets

X = I said hi to the manager

Y<t = Ho salutato ___



Source and Target Attributions
↓: Attributed →: Generated

	_Ho	_saluta	to	_la → _il	_manager	</s>
_I	0.0	0.0	0.0	0.06	0.0	0.0
_said	0.0	0.0	0.0	0.077	0.0	0.0
_hi	0.0	0.0	0.0	0.089	0.0	0.0
_to	0.0	0.0	0.0	0.042	0.0	0.0
_the	0.0	0.0	0.0	0.104	0.0	0.0
_manager	0.0	0.0	0.0	0.517	0.0	0.0
</s>	0.0	0.0	0.0	0.111	0.0	0.0
probability	0.44	0.844	0.916	0.785	0.331	0.26
contrast_prob	0.44	0.844	0.916	0.078	0.319	0.303
contrast_prob_diff	0.0	0.0	0.0	0.707	0.013	-0.043

Many step functions already available (Entropy, CE, CMI, PPL, KL-Div, …)

+ **custom step functions** registration to add yours!

Contrastive Attribution Inseq Tutorial

university of groningen | Dutch Research Council | INDEEP

Hands-On Notebook

university of groningen | Dutch Research Council | INDEEP

Examples on Google Colab

OR

https://tinyurl.com/indeep-masterclass-ex

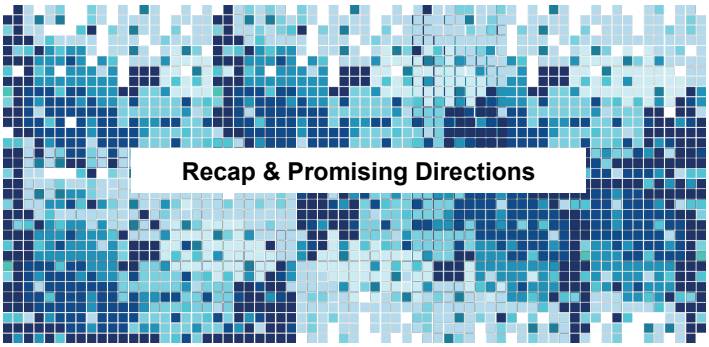
university of groningen | Dutch Research Council | INDEEP

Interested in Inseq? Let us know! 🎉

Contributions and feedback are more than welcome! 🤗

- Github - **inseq-team/inseq**
- Documentation - **inseq.readthedocs.io**
- Hugging Face Hub - **inseq**
- X (Twitter) - **@InseqDev**
- Discord - **see Github README**

university of groningen | Dutch Research Council | INDEEP



university of groningen | Dutch Research Council | INDEEP

After lunch: Let's discuss! 🗣️

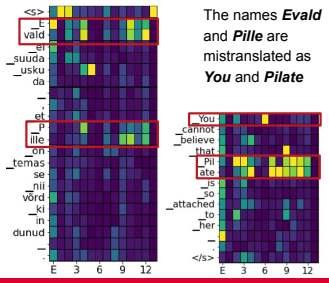
- Take a post-it, write on it:
  - your **name**
  - a **topic** that you'd like to discuss (e.g. task of interest, use case, an obstacle preventing you from using feature attribution in your work, a wish from research …)
- Paste it on the whiteboard
- After lunch we'll discuss in **groups**
- **Now:** Some success stories and promising directions to inspire you!

university of groningen | Dutch Research Council | INDEEP

Example: Coarse-to-Fine-grained Quality Estimation

Goal: Identify **word-level error spans** from sentence-level neural metrics (Fomicheva et al. 2022)

- Useful for obtaining **data-driven granular quality information**.
- Applications for human preference prediction currently under-explored (e.g. granular explanations for **reward model predictions** in RLHF)



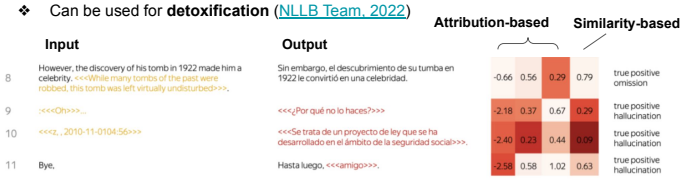
The names ***Evald*** and ***Pille*** are mistranslated as ***You*** and ***Pilate***

university of groningen | Dutch Research Council | INDEEP

Example: Hallucination Detection

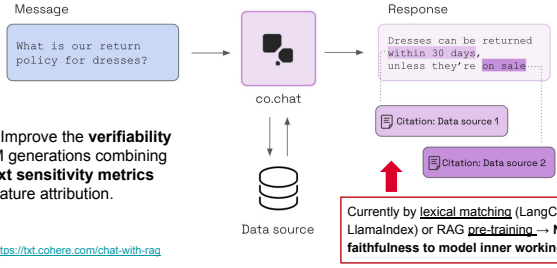
Goal: Detect hallucinations and omissions from context usage metrics (Dale et al. 2023)

- No additional training/modules, detect issues from **model internals** only.
- Can be used for **detoxification** (NLLB Team, 2022)



university of groningen | Dutch Research Council | INDEEP

Example: Trustworthy Retrieval-Augmented LLMs



Goal: Improve the **verifiability** of LLM generations combining **context sensitivity metrics** and feature attribution.

Currently by lexical matching (LangChain, LlamaIndex) or RAG pre-training → **No faithfulness to model inner workings!**

Source: https://txt.cohere.com/chat-with-rag

university of groningen | Dutch Research Council | INDEEP

PECoRe

Plausibility Evaluation for Context Reliance in LMs

(Most goods from Central America came here duty-free)

Target Context (FR): La plupart des biens d'Amérique centrale sont venus ici en franchise.

Source (EN): Yet eighty percent of the goods were taxed in Central American countries.

Contextless target (FR): Pourtant, 80% de nos marchandises ont été *taxées* dans les pays d'Amérique Centrale

Contextual target (FR): Pourtant, 80% de nos biens ont été *taxés* dans les pays d'Amérique centrale

Context-sensitive target and contextual cues (found by PECoRe) and contextless variant of the corresponding token. *Other changes* are not found to be context-dependent.

→ A new **data-driven method** to debug (im)plausible context dependence in LMs! Sarti et al.,2023

Thank you for your attention!

university of groningen | Dutch Research Council | INDEEP

@AriannaBisazza & @gsarti_

https://www.cs.rug.nl/~bisazza/ & gsarti.com

a.bisazza@rug.nl & g.sarti@rug.nl

Jelle Zuidema (UvA) discusses the promises and pitfalls for Explainability in the Generative AI boom, while connecting it to the work done by the hands-on groups

After the Masterclass, InDeep teamed up with Amsterdam AI for a Meetup on Alternatives for ChatGPT: *How good are open source and in-house LLMs?*

Harm de Vries (ServiceNow) presented the Big Code project, the pioneering project to create an open source dataset and generative AI model for generating code. He discussed details of the effort to determine the optimal dataset and model size, legal and ethical concerns about the dataset

Matthieu Laneuville (SURF) discussed the motivation and plans for an open source, community-driven Large Language Model for Dutch: GPT-NL

BigCode

Open and responsible development of Large Language Models for code

Harm de Vries | Staff Research Scientist @ ServiceNow | @harmdevries77

Today's talk

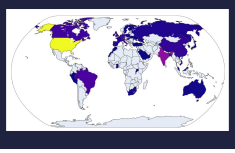
1. The BigCode Community
2. The Stack
3. StarCoder

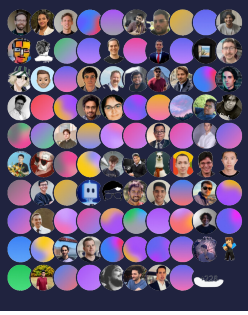
1. The BigCode community

BigCode: open-scientific collaboration

We are building LLMs for code in a collaborative way:

- 1100+ members
- 60+ countries





Developing Code LLMs is not only a technical problem!

- Consent of data subjects
  - Can you use third-party copyrighted data to train machine learning models?
- Privacy risks
  - Do LLMs distribute personally identifiable information without consent?
- Software safety and security
  - Code LLMs may be used to generate malware or may provide code suggestions that are less secure

Scraped public data sources

More info on the Governance Card

Closed development of LLMs

- Training data is not disclosed:
  - Content creators don't know if their data is used and there's no way to remove it
  - Limits scientific reproducibility
  - Potential benchmark contamination
- Model only available through API, which limits research on:
  - Safety and alignment
  - The model's inner workings (i.e. representations)
  - Adaptation methods like LoRA, and continuous prompt-tuning

Open & Responsible Research on LLMs

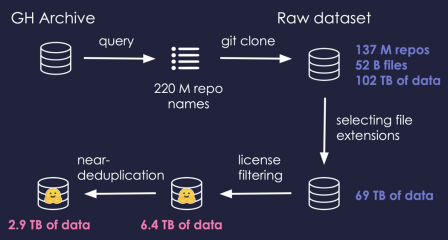
Open-access datasets	Open-access models	Reproducible research	Documentation
Permissive licensing	Model weights available	Data preprocessing scripts	Dataset cards
Data inspection	Fine-tuning scripts	Model training framework	Model cards
Opt-out available	Low-precision inference	R&D notebooks	Governance card
PII removal		Evaluation Harness	Intellectual property
Attribution			Code of conduct
			OpenRAIL licenses

2. The Stack

The Stack

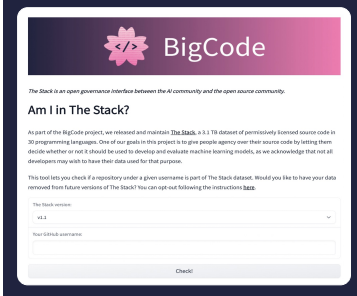
A dataset with **6.4TB** of **permissively licensed** source code in **358 programming languages** with a **data inspection** tool and **opt-out** mechanism

Data Collection



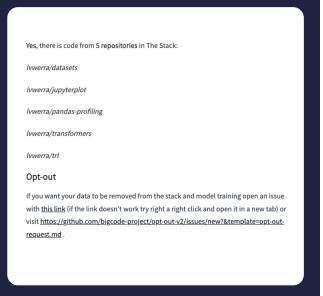
Find the filtered and deduplicated datasets at: www.hf.co/bigcode

Am I In The Stack?



https://huggingface.co/spaces/bigcode/in-the-stack

Opt-out



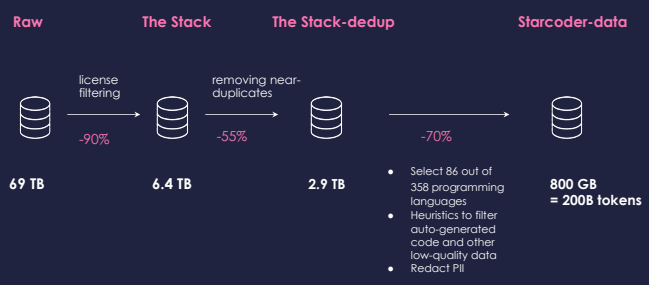
Feedback from the opt-out form

- *"It should be opt-in instead of opt-out"*
- *"It is unfair to use my code without compensation"*
- *"There's PII in my code and I don't want it to be publicly exposed"*
- *"My code is of poor quality and unsuitable for training your AI model"*
- *"I am not confident about the current state of AI code generation. I am concerned that the generated code could be traced back to me and I'm held liable for issues in that code."*

Jennifer Ding's community research: it's both **better to know** AND **better to have a choice**.

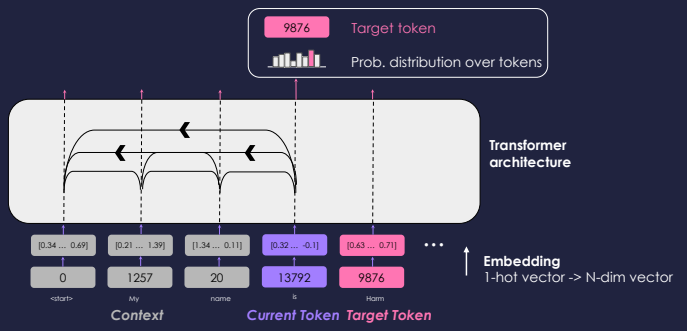
StarCoder

Starcoder training data



- Select 86 out of 358 programming languages
- Heuristics to filter auto-generated code and other low-quality data
- Redact PII

Generative Pretrained Transformer (GPT)



Scaling laws

The loss of language models follows a power law with respect to the number of parameters and training data

$$L(N, D) = E + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}}$$

→ *For a given compute budget, you can derive the optimal number of parameters and training tokens*

Kaplan scaling laws
model 5.5x, data 1.8x

GPT-3
175B params, 300B tokens

Chinchilla scaling laws
model 3.2x, data 3.2x

Chinchilla
65B params, 1.6T tokens

Compute-optimal ≠ optimal

LLaMA
7B params, 1T tokens

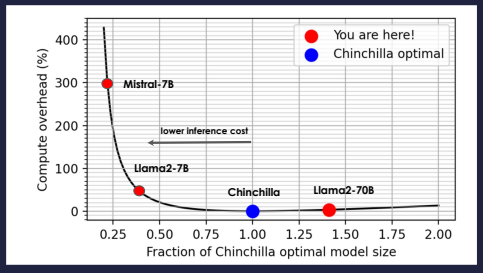
Jan 2020 — **Mar 2022** — **Feb 2023**

https://www.harmdevries.com/post/model-size-vs-compute-overhead/

"Harm's law"

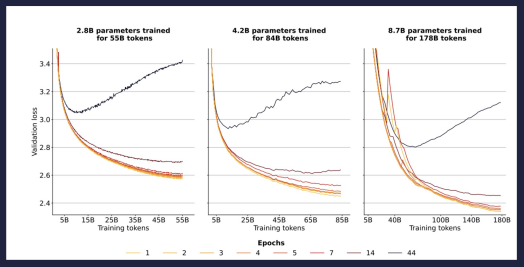
How much smaller can we make the compute-optimal model?
How much longer do we need to train? What's the compute-overhead?

$$L(k_N N, k_D D) = L(N, D)$$



https://huggingface.co/spaces/lvwerra/harms-law

How often can we repeat training data?

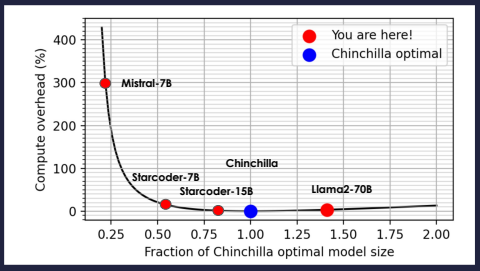


No significant degradation up to 5 epochs!

[Scaling data-constrained language models, Muennighoff et al]

Starcoder model size

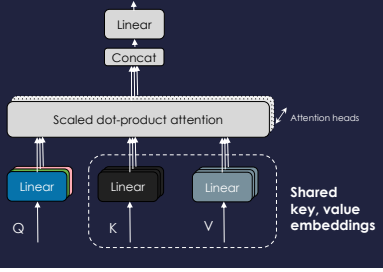
200B tokens x 5 epochs = 1 trillion training tokens



https://huggingface.co/spaces/lvwerra/harms-law

Model architecture

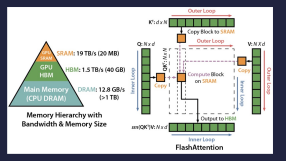
Multi-query attention



Shared key, value embeddings

8K context

thanks to FlashAttention



Training setup

Infrastructure: 512 A100 GPUs

Model Distribution: TP=4, PP=4, DP=32

Batch size: 4M tokens

(or 512 at 8,192 sequence length)

Training length: 1T tokens / 250k steps

Training time: 24 days

Tool: Megatron-LM (w/ MQA + FlashAttn)

(https://github.com/bigcode-project/Megatron-LM)

"smooth sailing"

Python Evaluation: HumanEval

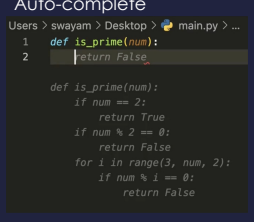
```
from typing import List

def filter_by_substring(strings: List[str],
substring: str) -> List[str]:

""" Filter an input list of strings only for
ones that contain given substring >>>
filter_by_substring([], 'a') [] >>>
filter_by_substring(['abc', 'bacd', 'cde',
'array'], 'a') ['abc', 'bacd', 'array'] """

Model generates function body

assert candidate(['xxx', 'asd', 'xxy', 'john
doe', 'xxxAAA', 'xxx'], 'xxx') == ['xxx',
'xxxAAA', 'xxx']
```

Model	Size	HumanEval pass@1
Open-access		
SantaCoder-1B	1B	18.1
DeciCoder-1B	1B	19.3
Replit-3B	3B	20.1
StableCode-3B	3B	20.2
StarCoderBase-3B	3B	21.5
StarCoderBase-7B	7B	28.4
CodeGen-Mono	16B	29.3
LLaMA-2	70B	29.9
CodeGen-2.5-Mono	7B	33.1
CodeGeeX-2	6B	33.5
StarCoder-15B	15B	33.6
OctoCoder*	15B	45.3
WizardCoder*	15B	58.1
Closed-access		
LaMDA	137B	14.0
PaLM	540B	26.2
code-cushman-001	12B	33.5
PaLM 2-S	N/A	37.6
code-davinci-002	175B	45.9
GPT-3.5	N/A	48.1
PanGu-Coder 2*	15B	61.6
GPT-4*	N/A	67.0

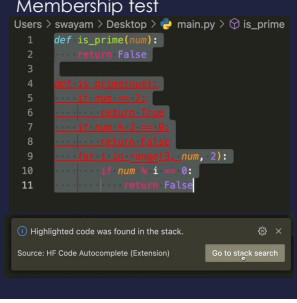
Code llama-34B: 48.8%

VSCode extension

Auto-complete

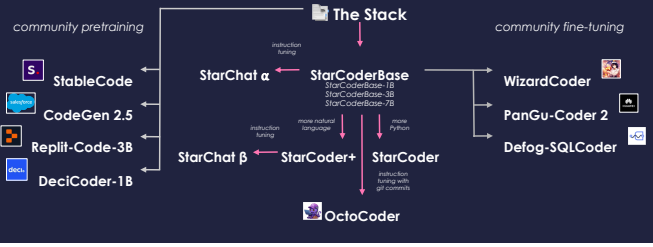


Membership test



https://marketplace.visualstudio.com/items?itemName=HuggingFace.huggingface-vscode

BigCode Ecosystem



Closing thoughts on open-source AI

- Frontier labs: OpenAI, Anthropic, Google, etc
- Allies of open-source AI: Meta, Mistral, Hugging Face, ServiceNow, etc
- Challenges for open-source AI to catch-up on frontier labs
  - depends on the support of large tech companies for:
    - compute
    - scarce AI talent
  - lack of user interaction data - openAI can use all the interactions with ChatGPT
- On the flipside, you might not need frontier models for the majority of use cases.
  - Many valuable enterprise use cases can be build on top of smaller open-source LLMs
- AI regulation
  - Copyright lawsuits
  - EU AI act
  - Biden's executive order
  - Regulation is the friend of the incumbent - Bill Gurley

Thank you! And come join us!

Questions?

www.bigcode-project.org

hf.co/bigcode

Building a truly open Dutch LLM

Matthieu Laneuville (SURF) &
Saskia Lensink (TNO)

November 2, 2023

Open?

Project	Availability						Documentation				Access	
(maker, base, URL)	Open code	LLM data	LLM weights	RLHF data	RLHF weights	License	Code	Architecture Preprint	Paper	Modelcard	Datasheet	Package	API
BLOOMZ	✓	✓	✓	✓	–	–	✓	✓	✓	✗	✓	✓	✗	✓
Open Assistant	✓	✓	✓	✓	✗	✓	✓	✓	–	✗	✗	✗	✓	✓
Pythia-Chat-Base-7...	✓	✓	✓	✓	✗	✓	✓	✓	–	✗	–	–	✓	✗
RedPajama-INCITE...	–	✓	✓	✓	✓	–	–	–	✗	✗	✓	✓	✗	–
dolly	✓	✓	✓	✓	✗	✓	✓	✓	–	✗	✗	✗	✓	✗
OpenChat	✓	✗	–	✗	✓	–	–	✓	✓	✗	–	✗	✓	–
MPT-7B Instruct	✓	–	✓	–	✗	✓	✓	–	✗	✗	✓	✗	✓	✗
trlx	✓	✓	✓	–	✗	✓	✓	–	✗	✗	✗	✗	–	✓
MPT-30B Instruct	✓	–	✓	–	✗	✓	✓	–	✗	✗	–	✗	✓	–
Vicuna 13B v 1.3	✓	–	✓	✗	✗	–	✓	✗	✓	✗	–	✗	✓	–
minChatGPT	✓	✓	✓	–	✗	✓	✓	–	✗	✗	✗	✗	✗	✓
ChatRWKV	✓	–	✓	✗	✗	✓	–	–	–	✗	✗	✗	✓	–
Cerebras-GPT-111M	✓	✓	✓	✓	✗	✓	✗	✓	–	✗	✗	✗	✗	✗
WizardLM 13B v1.2	–	✗	–	✓	✓	–	–	✓	✓	✗	✗	✗	✗	✗

https://opening-up-chatgpt.github.io/

Dutch?

https://medium.com/@socialcreature/ai-and-the-american-smile-76d23a0fbfaf

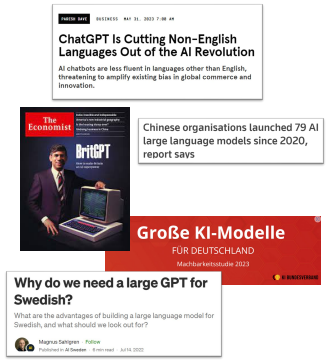
Dutch?

https://medium.com/@socialcreature/ai-and-the-american-smile-76d23a0fbfaf

Why a Dutch LLM?

A major obstacle to the use of AI in the Netherlands is that existing algorithms are not properly trained in the Dutch language. [...] This problem [...] applies to the entire public sector and to all Dutch-language interactions in the market.
-- https://nlaic.com/en/use-case/nain-dutch-ai-for-dutch-language/

- Risks around bias, inclusion and explainability are not sufficiently guaranteed in current solutions
- Existing models of tech giants do not work sufficiently or cannot be used due to IP&C concerns
  - What if we want to use those models in a medical context? In an education context?
- Also, developing skills & capacity in the public sector helps with long term autonomy



The GPT-NL ambition

- A modern Dutch model

The GPT-NL ambition

- Develop, strengthen and consolidate **digital sovereignty** through our own Dutch language model.
- Increase **strategic autonomy** in the underlying knowledge and technology.
- **Facilitate short-cycle, application-oriented research** into the use of language models in the broad context of social challenges.
- Enable **fundamental research into values such as bias and transparency** in large language models.

WRR. Mission AI report.

COMMENT | 19 October 2023

Living guidelines for generative AI — why scientists must oversee its use

Establish an independent scientific body to test and certify generative artificial intelligence, before the technology damages science and public trust.

Claudi L. Bockting, Eva A. M. van Dis, Robert van Rooij, Willem Zuidema & Johan Bollen

In practice…

Financing
Initial subsidy from Tender 'dream facilities' ministry of Economic Affairs, ~14 M€
Public-private partnerships and joint innovation centers to exploit the facility
Not-for-profit approach to cost structure (self-sustaining)

Facility
A Dutch foundation model, facilitating further research, fine-tuning and deployment
Trained and deployed on hardware hosted by SURF, new hardware will be purchased (88+16 H100)
A factor 1,000 times larger than the current Dutch language models
Both dataset and model open for inspection and research

Data
Existing open-source data sets (think of The Pile, Wikipedia, etc.) are a possibility
Text data extracted from high-quality sources such as government websites, forums, etc.
Public data (e.g., CLARIAH and CLARIN-NL, the Royal Library, etc.)
Data curation will be absolutely crucial, and choices will have to be made
The dataset used to train the base model will be made available under CC-BY-4.0

In practice…

Potential working groups (inspired by BigScience/BLOOM)

- Data
- Engineering / scaling – focus on the technical challenges of training (scaling, specialized hardware)
- Evaluation
- Legal & ethical issues
- Modelling
- Model & data sharing – prepare ethical framework and tools for sharing
- Governance / workshop organization – main organizing / coordination group
- Carbon footprint / sustainability – what is the carbon footprint from the project and how to reduce it
- Challenges for application X – series of working groups focused on specific applications

	jaar 1												jaar 2			
	jan			apr			jul			okt		dec	jan			apr
Dataset opbouw																
Inrichting ontwikkel cluster																
Algoritme ontwikkeling																
Training van taalmodel																
Inrichting operationeel cluster																
Platform ontwikkeling																
Inrichting mobiel bijtrain cluster																

Figuur 1: Activiteiten en milestones door de tijd

Opening remarks

- TNO/SURF led community owned
- How do we facilitate responsible experimentation together?
- How do we embed this in the larger international context?
- What should be our shared standard for openness?

Rijksoverheid

Home › Actueel › Nieuws ›

14 Nederlandse onderzoeksfaciliteiten krijgen 184 miljoen euro ondersteuning

Nieuwsbericht | 02-11-2023 | 12:36

Jelle Zuidema discusses and shares with **Konstantinos Papakostas** (ZetaAlpha), **Iva Gornishka** (City of Amsterdam), **Jelke Bloem** (ILLC, University of Amsterdam) and **Oskar van der Wal** (ILLC, University of Amsterdam) their perspectives on how to evaluate such models and make sure they are useful, safe and reliable

Thank you for your interest in the **InDeep Project**

For more information about the project, members, research and our upcoming events, visit:
projects.illc.uva.nl/indeep/

nationale wetenschaps agenda

NWO

InDeep

INTERPRETING DEEP LEARNING MODELS FOR TEXT AND SOUND

Radboud University

University of Amsterdam

Tilburg University

VU Vrije Universiteit Amsterdam

university of groningen

OUR PARTNERS:

KPN - IDFUSE - TNO - TEXTKERNEL - CHORDIFY - GLOBAL TEXTWARE
DELOITTE - FLOODTAGS