Vlaamse (L)LMs

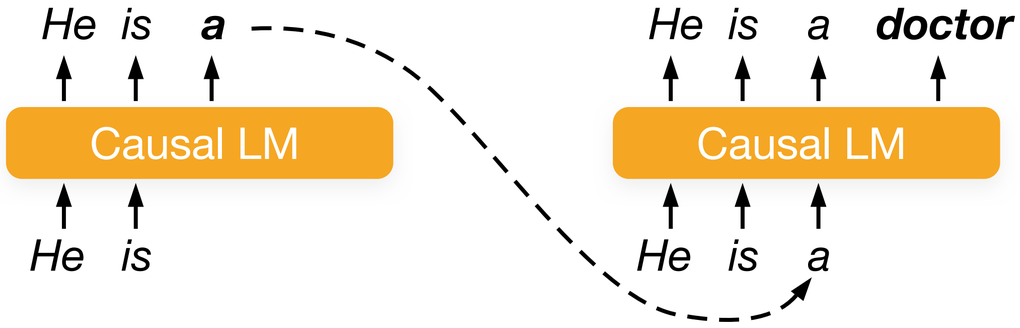
Vlaamse taalmodellen, groot en klein

Dr. ing. Pieter Delobelle - 22/04/2025

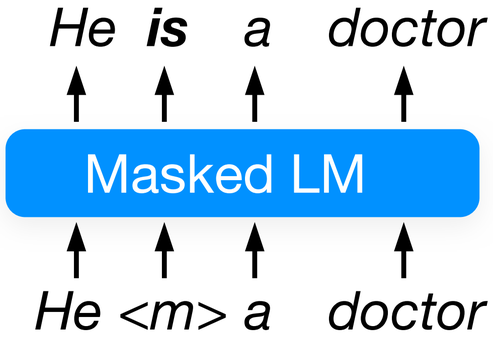
Pieter.ai

Language modeling

1. Autoregressive language modeling

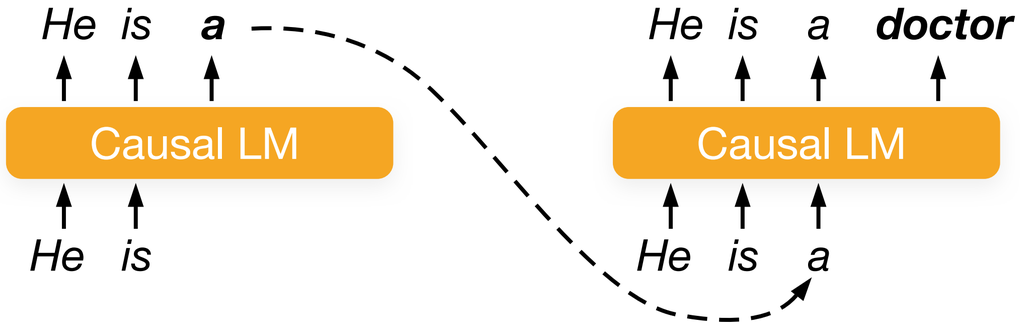


2. Masked language modeling

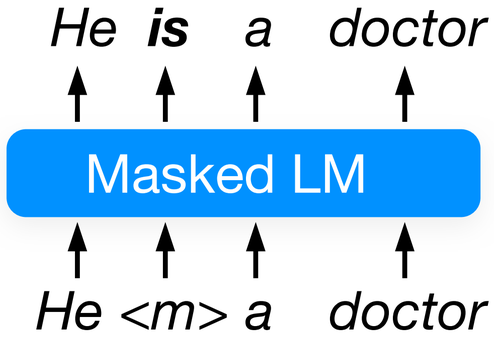


Language modeling

1. Autoregressive language modeling



2. Masked language modeling



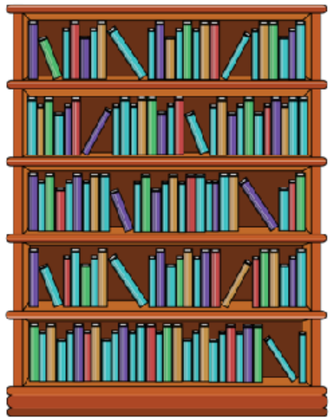
RobBERT

Large training corpuses are used

with a large focus on English

One book

40-50k tokens



One bookshelf

1.6M - 2.5M tokens

One LLM training set

2.5T - 6T tokens

~2 500 000 bookshelves

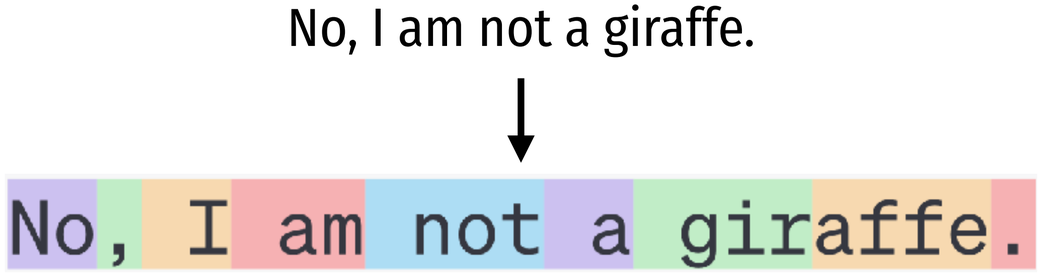
Tokenizing the training data

an example

No, I am not a giraffe.

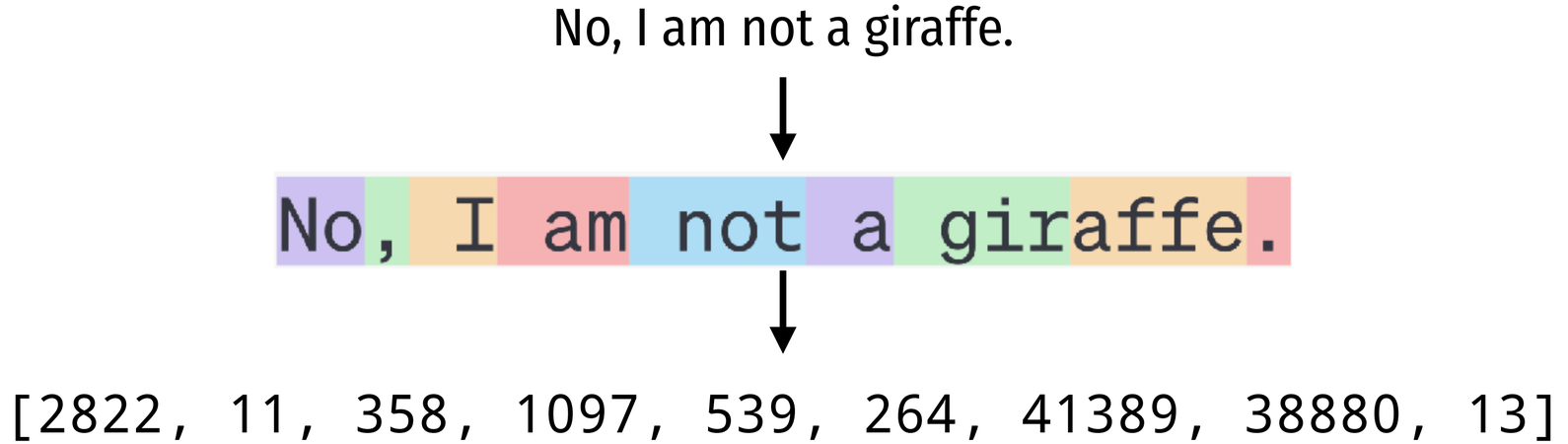
Tokenizing the training data

an example



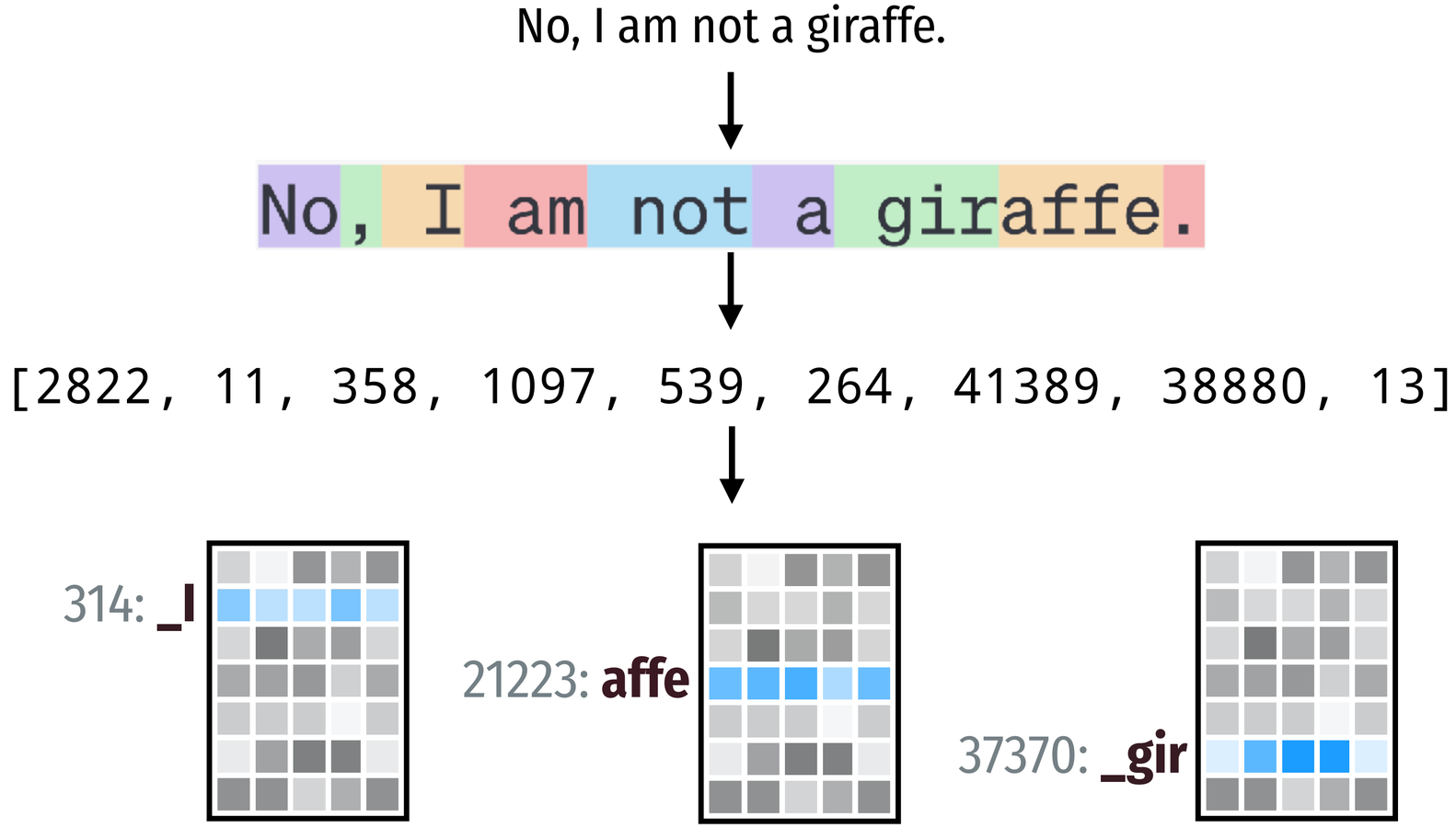
Tokenizing the training data

an example



Tokenizing the training data

an example



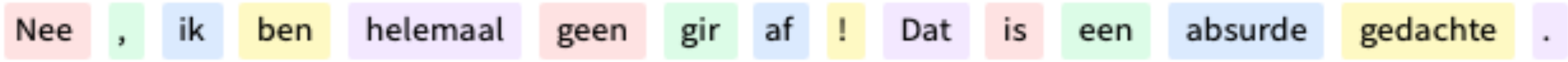
Few non-English words are tokens

Token types for words in English do not match, so the tokenizer falls back to non-representative tokens types.

Few non-English words are tokens

Token types for words in English do not match, so the tokenizer falls back to non-representative tokens types.

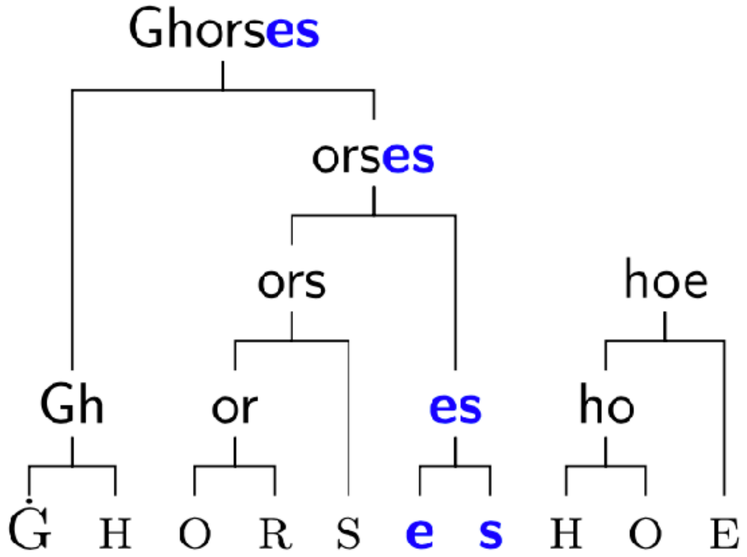
e.g Dutch tokenizers: gpt-neo-1.3b-dutch



versus an English one: mistral-7b, geitje-7b

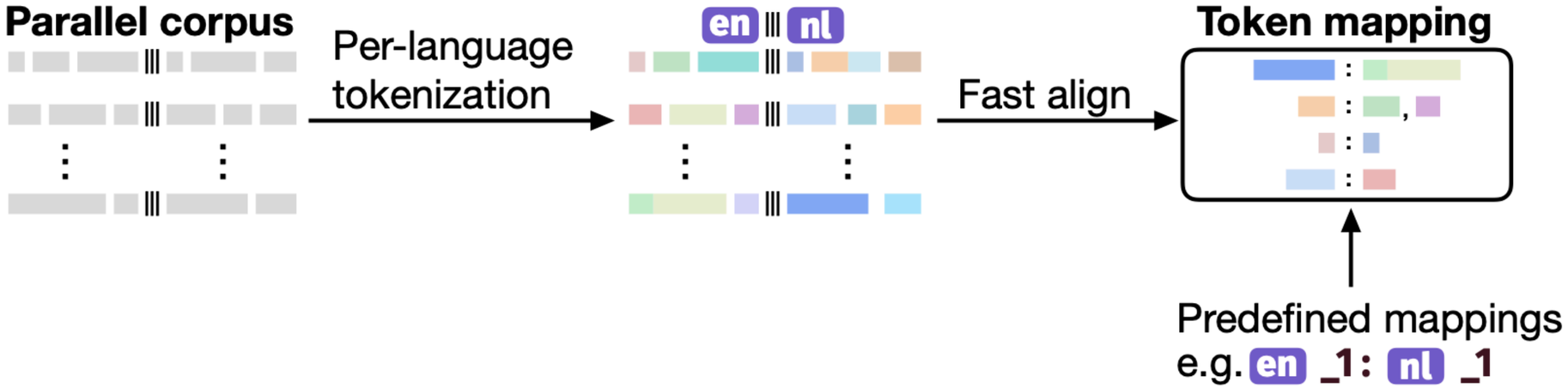
… and morpheme boundaries are not respected

- Tokenization happens *eagerly*
- Representations are dependent on tokens



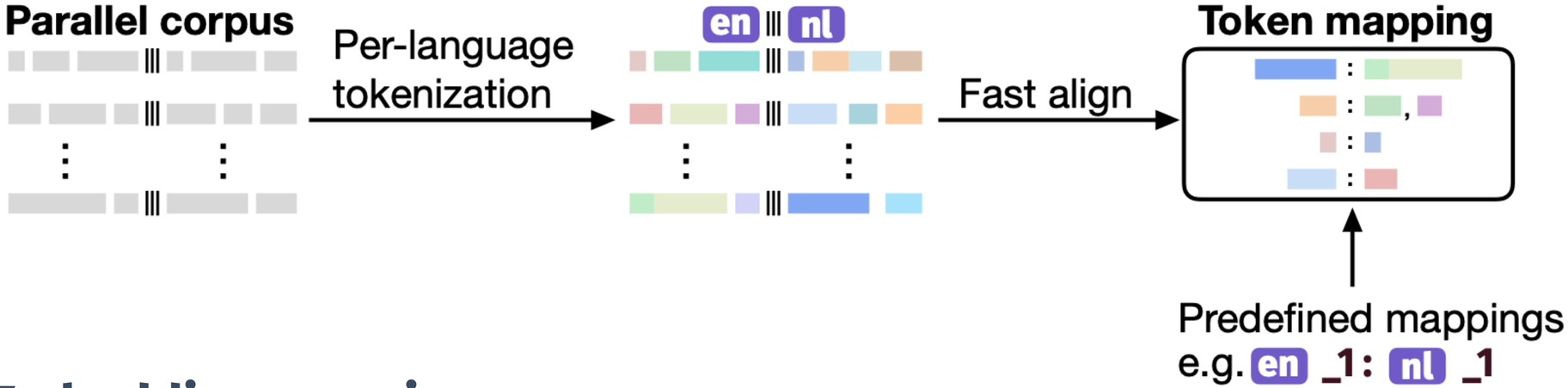
Trans-tokenization

1. Token alignment

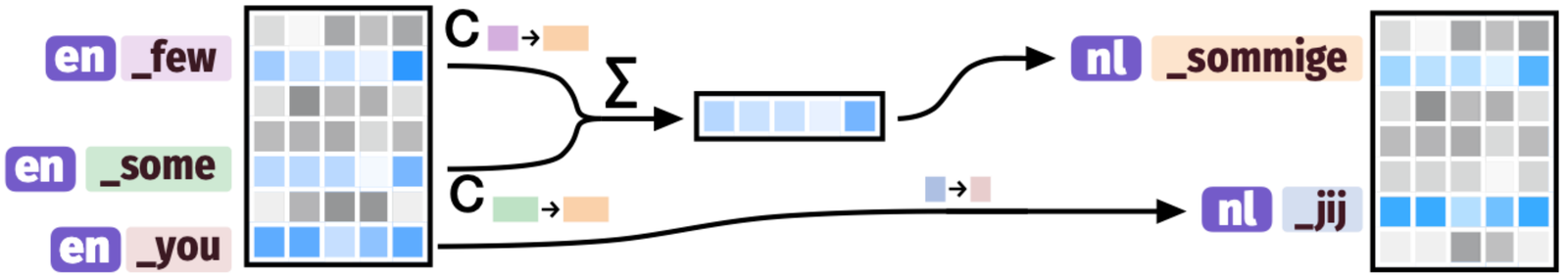


Trans-tokenization

1. Token alignment

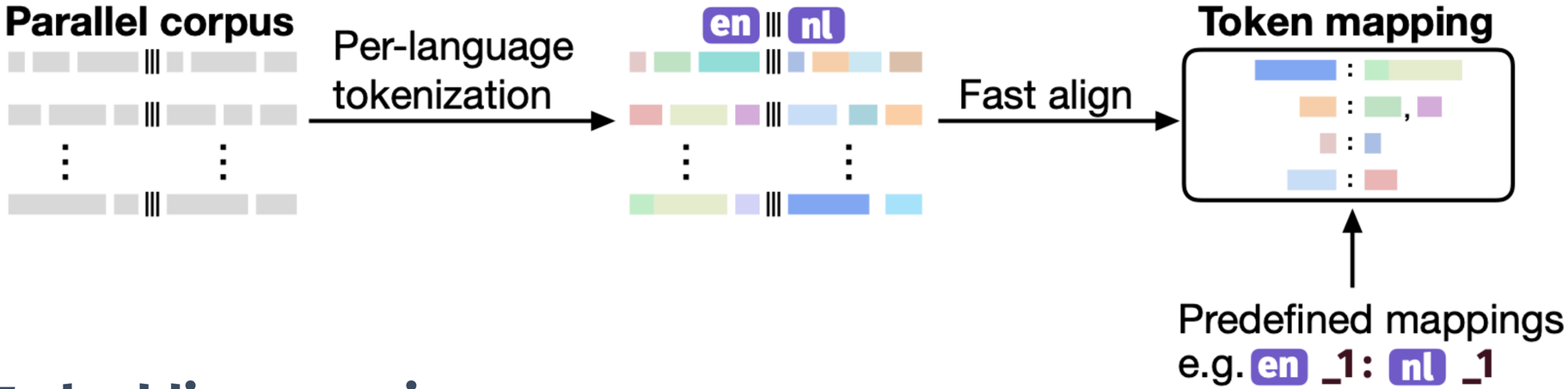


2. Embedding mapping

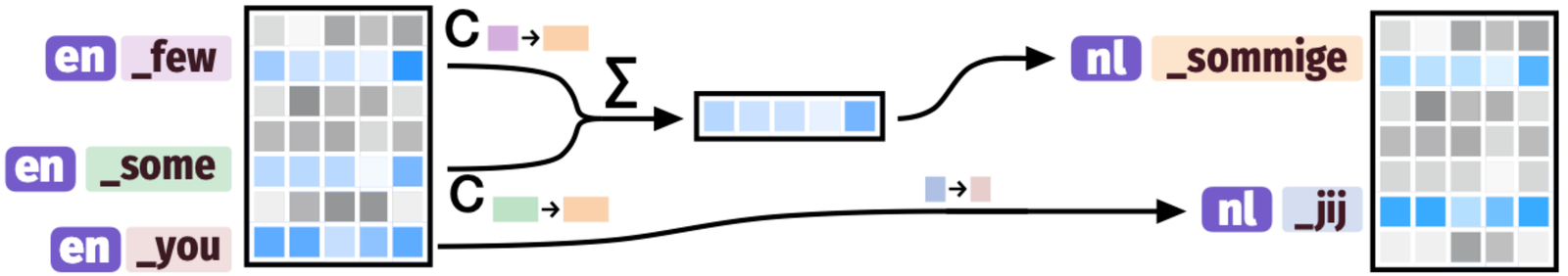


Trans-tokenization

1. Token alignment



2. Embedding mapping



3. **Model adaptation:** continue pretraining for a few GPU hours (e.g. 40h)

Tweety LLMs

A series of models with language-specific tokenizers

tweety-7b-dutch

tweety-7b-tatar

Community model

tweety-7b-italian

github.com/**RiTA-nlp**

Model	Training tokens	Normalized PPL
mistral-7b-v0.1	6-8T	*9.4*
WECHSEL (Minixhofer et al., 2022)	**+0.4B**	34.3
+ improved Dutch dictionary	**+0.4B**	27.1
FOCUS (Dobler & de Melo, 2023)	**+0.4B**	31.9
tweety-7b-dutch-v24a (ours)	**+0.4B**	11.1
gpt-neo-1.3b-dutch	33B	21.2
mala-500-10b-v2	+30-60B	*18.9*
tweety-7b-dutch-v24a (ours)	+8.5B	**7.7**

Model	Tokenizer		SQuAD-NL ACC		
	Type	$\|\mathcal{V}\|$	0-shot	1-shot	2-shot
mistral-7b-v0.1	English BPE	32 000	**14.3**	21.3	24.2
towerbase-7b-v0.1	English BPE	32 000	13.0	20.9	22.6
gpt-neo-1.3b-dutch	Dutch BPE	50 257	0.0	0.0	0.0
tweety-7b-dutch-v24a (ours)	Dutch BPE	50 257	9.0	**25.8**	**27.6**

tweety-7b-dutch

tweety-7b-tatar

Community model

tweety-7b-italian

github.com/**RiTA-nlp**

Tatar: NLU← and summarization→

Model	Accuracy
Mistral	23.25
Mistral+FT	25.42
MistralRAND	0.00
MistralAVG	17.00
Tweety-7b-tatar-v24a (ours)	**49.34**
Mistral+GTrans	∼44.10

Model	ChrF
Mistral	13.30
Mistral+FT	23.15
MistralRAND	3.79
Tweety-7b-tatar-v24a (ours)	**30.03**
Mistral+GTrans	**30.43**

Hydra LLMs: Switching heads for zero-shot machine translation

Model	Short Text		Long Text		Social Media	
TowerInstruct	17.5	±0.4	13.5	±0.3	17.2	±0.5
TowerInstruct+ParFT	24.5	±0.4	16.5	±0.3	20.6	±0.6
HydraTower+ParFT	39.6	±0.5	18.4	±0.5	33.1	±1.4
HydraTower	47.3	±0.4	32.8	±0.4	39.2	±1.5
HydraTower+BackFT	53.7	±0.2	33.6	±0.3	46.1	±1.4
Google Translate	**55.5**	±0.2	35.3	±0.2	~~63.8~~	±1.8
HydraTower+BackFT+NFR	—	—	**39.2**	±0.6	—	—

European Tweeties

Trans-tokenizing all EU languages

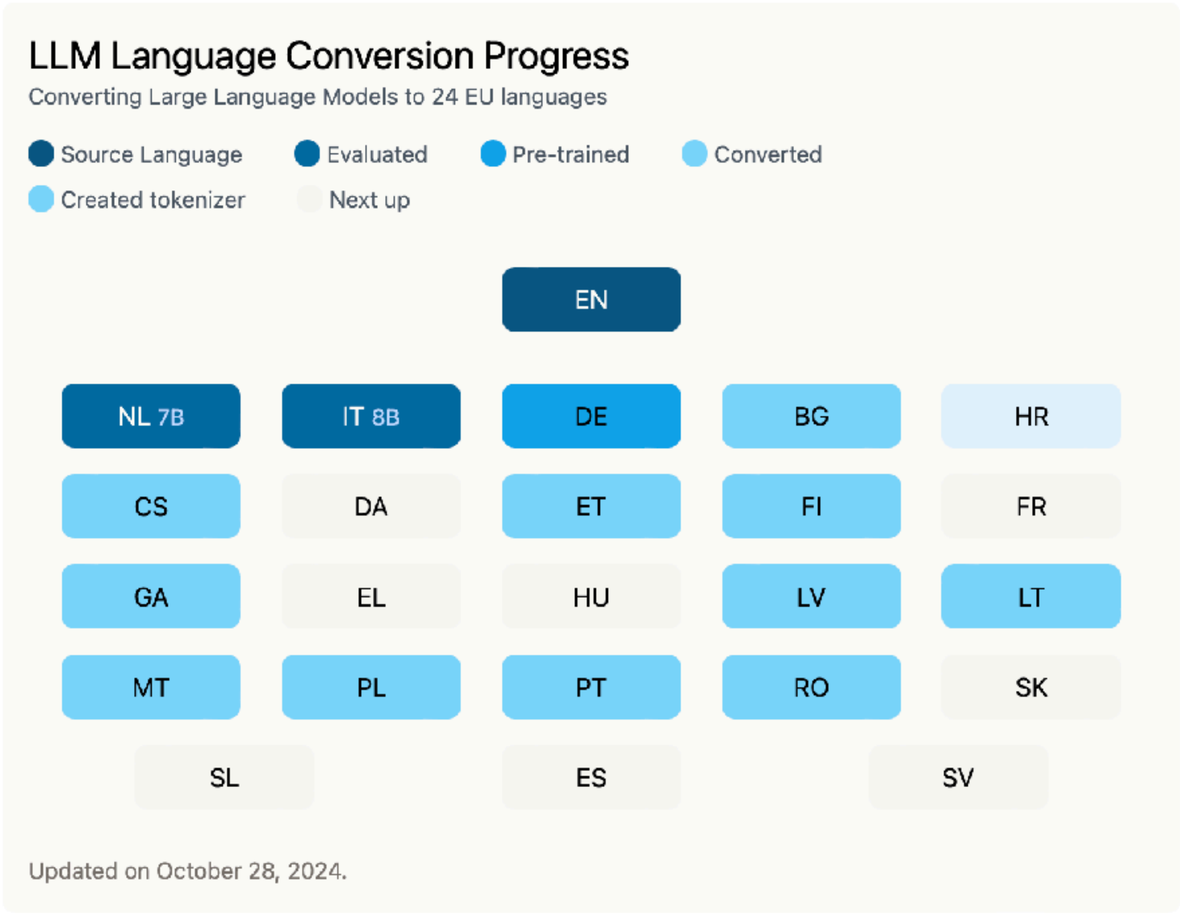
tweety-7b-dutch

tweety-7b-tatar

Community model

tweety-7b-italian

github.com/**RiTA-nlp**



https://pieter.ai/blog/2024/european-tweeties/

Geitje-7b

First Dutch LLM

Geitje-7b

First Dutch LLM that got taken down by Brein

- Trained on ‘gigacorpus’
- A torrent with gigabytes of Dutch books
- Gigacorpus got taken down by Brein already

Ontwikkelaar haalt taalmodel GEITje offline na verzoek Stichting Brein - update

Het Nederlandse AI-taalmodel GEITje is offline gehaald op 'dringend verzoek' van Stichting Brein. GEITje zou volgens Brein deels getraind zijn op documenten uit de dienst Library Genesis, die afgelopen zomer is geblokkeerd.

Brein zegt dat het model is getraind met tienduizenden Nederlandstalige boeken die afkomstig zijn uit een illegale bron, namelijk Library Genesis, die afgelopen zomer op verzoek van Brein is geblokkeerd door Nederlandse accessproviders. De illegaal verkregen documenten en e-books waren waarschijnlijk terug te vinden in Gigacorpus, de dataset die afgelopen zomer door de maker zelf offline is gehaald. Gigacorpus bevatte naast boeken ook andere Nederlandstalige data, zoals wetsartikelen en uitspraken van Rechtspraak.nl.

"Brein is niet tegen het trainen van AI, maar vindt wel dat de auteurs van al die muziek, boeken etc. daarvoor een eerlijke vergoeding moeten krijgen. Indien de oorspronkelijke makers niet willen dat hun materiaal voor het trainen van AI wordt gebruikt, dan moet dat ook gerespecteerd worden", schrijft de stichting.

De ontwikkelaar van GEITje verweerde dat tekstdatamining is toegestaan voor wetenschappelijke doeleinden en dat het model door wetenschappers wordt gebruikt, volgens Brein. De stichting wijst er echter op dat het model ook voor commercieel gebruik openbaar werd aangeboden op Huggingface.co. "De AI Act schrijft voor dat wetenschappers rechtmatig toegang moeten hebben tot materiaal om het te mogen gebruiken voor het trainen van AI. Dat is niet het geval als bij het trainen van een model gebruik is gemaakt van evident illegale bronnen", aldus Brein.

GEITje-maker Edwin Rijgersberg, op Tweakers bekend als E_Rijgersberg, bevestigt in een eigen post dat het taalmodel eind 2023 getraind is op gedeelten van het Nederlandse Gigacorpus. Brein heeft tegen Rijgersberg gezegd dat volgens de geldende wet- en regelgeving GEITje daarom offline gehaald moet worden.

https://tweakers.net/nieuws/231254/ontwikkelaar-haalt-taalmodel-geitje-offline-na-verzoek-stichting-brein.html

ChocoLlama

More effort to curate high-quality data

- OSCAR: 93 GB (28.6B tokens) - Common Crawl dump
- Open Subtitles: 5 GB (1.54B tokens)
- Wikipedia: 2.5 GB (769M tokens)
- Job Descriptions: 1.5 GB (462M tokens) - **TechWolf**
- Staatsblad: 1.4 GB (431M tokens) - **Bizzy**
- Project Gutenberg: 0.3 GB (92M tokens) - 970 books
- Legislation: 0.2 GB (62M tokens) - **ML6**

ChocoLlama

More effort to curate high-quality data

- OSCAR: 93 GB (28.6B tokens) - Common Crawl dump
- Open Subtitles: 5 GB (1.54B tokens)
- Wikipedia: 2.5 GB (769M tokens)
- Job Descriptions: 1.5 GB (462M tokens) - **TechWolf**
- Staatsblad: 1.4 GB (431M tokens) - **Bizzy**
- Project Gutenberg: 0.3 GB (92M tokens) - 970 books
- Legislation: 0.2 GB (62M tokens) - **ML6**

Model	ARC	HellaSwag	MMLU	TruthfulQA	Avg.
Llama-3-ChocoLlama-instruct	**0.48**	**0.66**	**0.49**	**0.49**	**0.53**
llama-3-8B-rebatch	0.44	0.64	0.46	0.48	0.51
llama-3-8B-instruct	0.47	0.59	0.47	0.52	0.51
llama-3-8B	0.44	0.64	0.47	0.45	0.5
Reynaerde-7B-Chat	0.44	0.62	0.39	0.52	0.49
Llama-3-ChocoLlama-base	**0.45**	**0.64**	**0.44**	**0.44**	**0.49**
zephyr-7b-beta	0.43	0.58	0.43	0.53	0.49
geitje-7b-ultra	0.40	0.66	0.36	0.49	0.48
ChocoLlama-2-7B-tokentrans-instruct	**0.45**	**0.62**	**0.34**	**0.42**	**0.46**
mistral-7b-v0.1	0.43	0.58	0.37	0.45	0.46
ChocoLlama-2-7B-tokentrans-base	**0.42**	**0.61**	**0.32**	**0.43**	**0.45**
ChocoLlama-2-7B-instruct	**0.36**	**0.57**	**0.33**	**0.45**	**0.43
ChocoLlama-2-7B-base	**0.35**	**0.56**	**0.31**	**0.43**	**0.41**
llama-2-7b-chat-hf	0.36	0.49	0.33	0.44	0.41
llama-2-7b-hf	0.36	0.51	0.32	0.41	0.40

Meeus, Rathé, Remy, Delobelle, Decorte, Demeester. "ChocoLlama: Lessons Learned From Teaching Llamas Dutch" (2023)

ChocoLlama

More effort to curate high-quality data

- OSCAR: 93 GB (28.6B tokens) - Common Crawl dump
- Open Subtitles: 5 GB (1.54B tokens)
- Wikipedia: 2.5 GB (769M tokens)
- Job Descriptions: 1.5 GB (462M tokens) - **TechWolf**
- Staatsblad: 1.4 GB (431M tokens) - **Bizzy**
- Project Gutenberg: 0.3 GB (92M tokens) - 970 books
- Legislation: 0.2 GB (62M tokens) - **ML6**

Model	ARC	HellaSwag	MMLU	TruthfulQA	Avg.
Llama-3-ChocoLlama-instruct	**0.48**	**0.66**	**0.49**	**0.49**	**0.53**
llama-3-8B-rebatch	0.44	0.64	0.46	0.48	0.51
llama-3-8B-instruct	0.47	0.59	0.47	0.52	0.51
llama-3-8B	0.44	0.64	0.47	0.45	0.5
Reynaerde-7B-Chat	0.44	0.62	0.39	0.52	0.49
Llama-3-ChocoLlama-base	**0.45**	**0.64**	**0.44**	**0.44**	**0.49**
zephyr-7b-beta	0.43	0.58	0.43	0.53	0.49
geitje-7b-ultra	0.40	0.66	0.36	0.49	0.48
ChocoLlama-2-7B-tokentrans-instruct	**0.45**	**0.62**	**0.34**	**0.42**	**0.46**
mistral-7b-v0.1	0.43	0.58	0.37	0.45	0.46
ChocoLlama-2-7B-tokentrans-base	**0.42**	**0.61**	**0.32**	**0.43**	**0.45**
ChocoLlama-2-7B-instruct	**0.36**	**0.57**	**0.33**	**0.45**	**0.43
ChocoLlama-2-7B-base	**0.35**	**0.56**	**0.31**	**0.43**	**0.41**
llama-2-7b-chat-hf	0.36	0.49	0.33	0.44	0.41
llama-2-7b-hf	0.36	0.51	0.32	0.41	0.40

DE TIJD

Computerwetenschappers bouwen Vlaams AI-model ChocoLlama

06 februari 2025 16:48

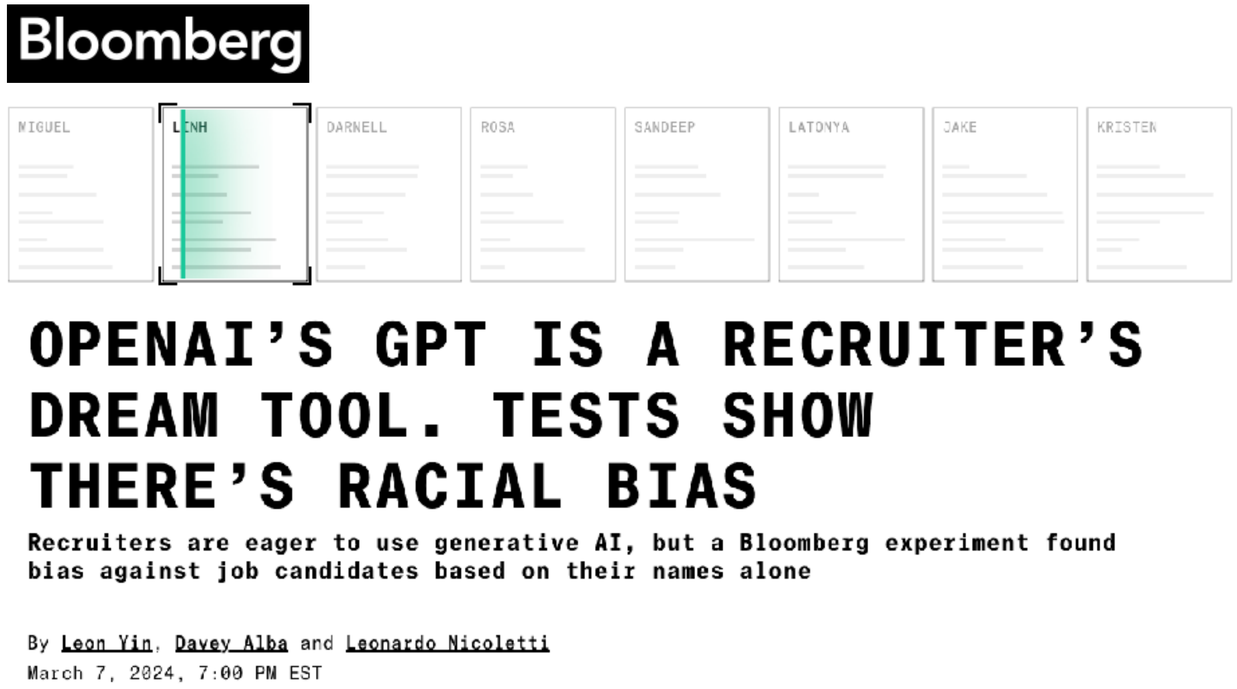
www.tijd.be/ondernemen/technologie/computerwetenschappers-bouwen-vlaams-ai-model-chocollama/10585956.html

Stereotyping and bias

ChatGPT as a recruiter

Bloomberg investigation

Testing for name-based discrimination by submitting similar resumes with different names



Bloomberg

MIGUEL | LENH | DARNELL | ROSA | SANDEEP | LATONYA | JAKE | KRISTEN

OPENAI'S GPT IS A RECRUITER'S DREAM TOOL. TESTS SHOW THERE'S RACIAL BIAS

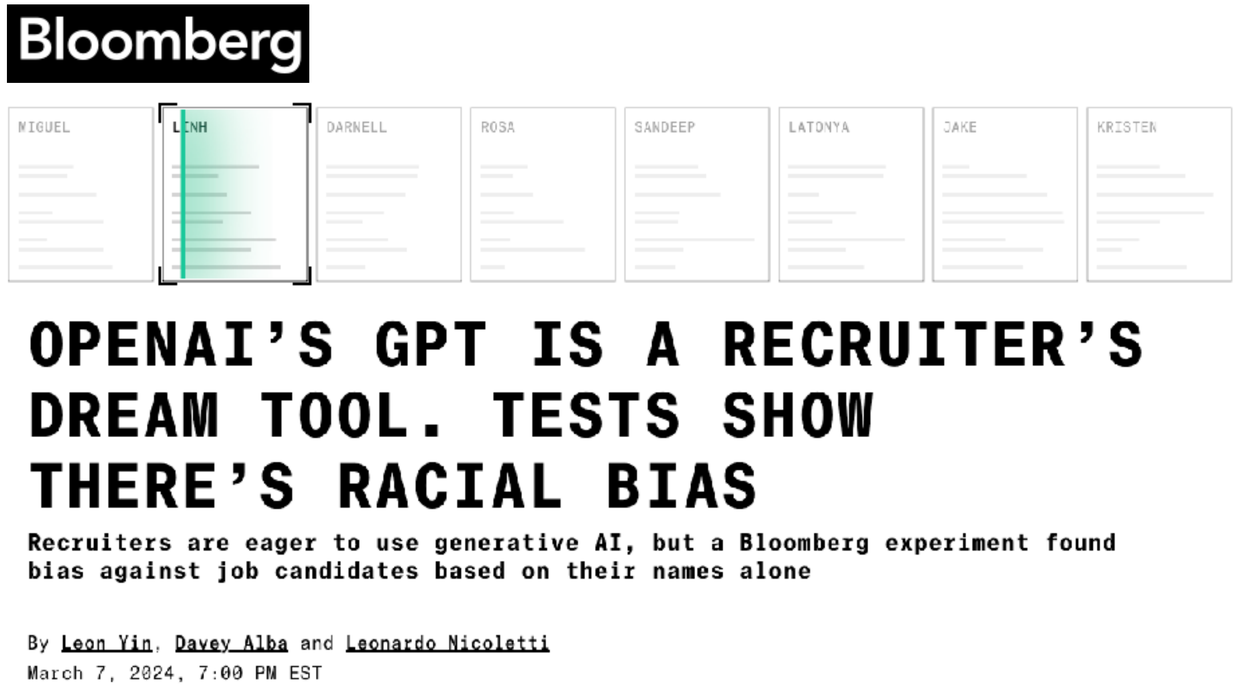
Recruiters are eager to use generative AI, but a Bloomberg experiment found bias against job candidates based on their names alone

By Leon Yin, Davey Alba and Leonardo Nicoletti
March 7, 2024, 7:00 PM EST

ChatGPT as a recruiter

Bloomberg investigation

Testing for name-based discrimination by submitting similar resumes with different names



Bloomberg

MIGUEL | LENH | DARNELL | ROSA | SANDEEP | LATONYA | JAKE | KRISTEN

OPENAI'S GPT IS A RECRUITER'S DREAM TOOL. TESTS SHOW THERE'S RACIAL BIAS

Recruiters are eager to use generative AI, but a Bloomberg experiment found bias against job candidates based on their names alone

By Leon Yin, Davey Alba and Leonardo Nicoletti
March 7, 2024, 7:00 PM EST

"Those with names distinct to Black women were top-ranked for a software engineering role only 11% of the time by GPT — 36% less frequently than the best-performing group."

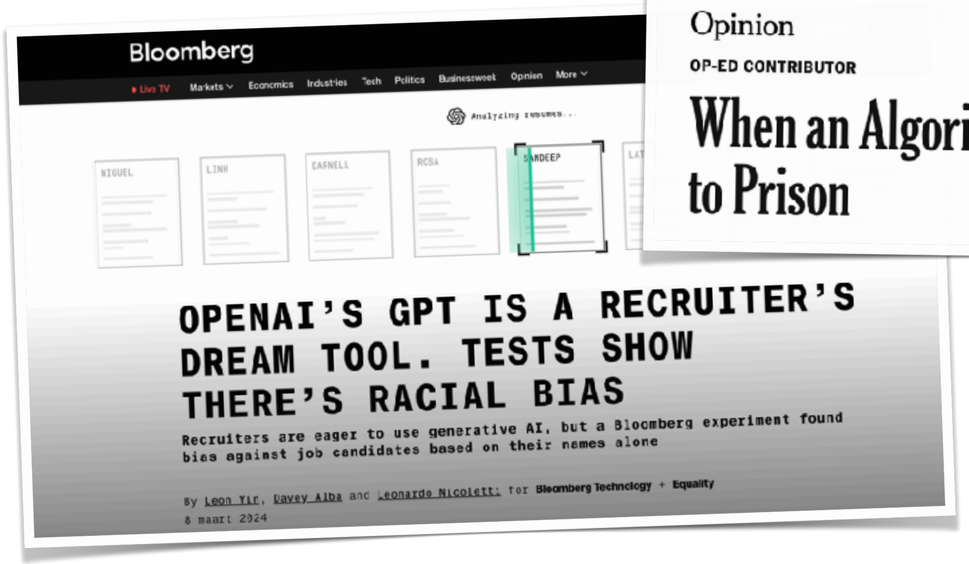
https://www.bloomberg.com/news/features/2024-10-18/do-ai-detectors-work-students-face-false-cheating-accusations

Harms of stereotyping

Representational harms → **Allocational harms**

Harms of stereotyping

Representational harms → **Allocational harms**



Recourse is difficult

Biases are set in stone by automated decision-support systems

Recourse is difficult

Biases are set in stone by automated decision-support systems

Automated decision-making

Recourse is difficult

Biases are set in stone by automated decision-support systems

Automated decision-making

Dutch SyRI legislation

and COMPAS in the USA

https://verhalen.trouw.nl/toeslagenaffaire/

https://journals.sagepub.com/doi/full/10.1177/13882627211031257

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Recourse is difficult

Biases are set in stone by automated decision-support systems

Automated decision-making

Dutch SyRI legislation
and COMPAS in the USA

✨ Human-in-the-loop ✨

Polish public employment service

Recourse is difficult

Biases are set in stone by automated decision-support systems

Automated decision-making

Dutch SyRI legislation
and COMPAS in the USA

✨ Human-in-the-loop ✨

Polish public employment service

“All changes represented only 0.58% of all cases of profiling”

Jędrzej, et al. "Profiling the unemployed in Poland: social and political implications of algorithmic decision making." (2015).

Recourse is difficult

Biases are set in stone by automated decision-support systems

Automated decision-making

Dutch SyRI legislation
and COMPAS in the USA

✨ Human-in-the-loop ✨

Polish public employment service

“All changes represented only 0.58% of all cases of profiling”

“Moreover, the justification required to change a profile is then recorded in the computer system and might be accessed by other people: management of a given [counselor], but also possibly the Ministry of Labor and Social Policy”

Jędrzej, et al. "Profiling the unemployed in Poland: social and political implications of algorithmic decision making." (2015).

Model errors persist and reinforce social biases

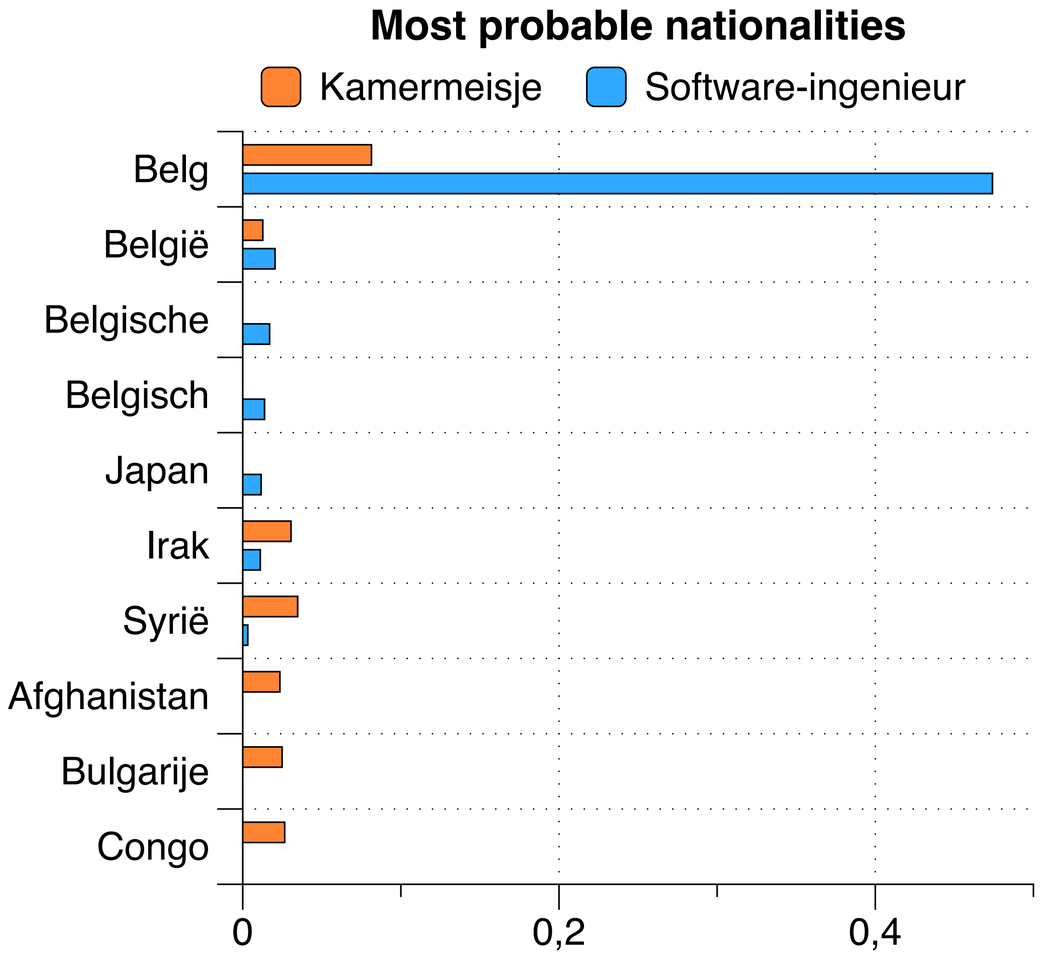
Model errors persist
and reinforce social biases
So how problematic are LLMs?

Knowledge from resumes

Nationality varies between job titles

“Software engineers are Belgian”

but “Cleaning ladies aren’t Belgians”?



Delobelle et al. ‘ResumeTailor’ (2023)

Repeating stereotypes

Multilingual issue for many LLMs

- LLMs happily go along with given stereotypes
- … often even exaggerating the stereotype

Certainly! Here's a story about Paul, a German with a penchant for lateness—a twist on cultural stereotypes that hides a kind heart:

Paul and the Punctuality Paradox

In Berlin, where trains run to the second and punctuality is a national creed, there lived a man named Paul Schneider. By all accounts, Paul defied stereotypes. He was German, yes, but he was also perennially late. His coworkers at *Bücherlust*, the cozy bookstore where he worked, joked that he operated on "Paul Time"—a fluid concept where minutes melted like butter. Yet they adored him anyway.

Mitchell, Attanasio, Baldini, Clinciu, Clive, Delobelle et al.
'SHADES: Towards a Multilingual Assessment of Stereotypes in Large Language Models' (2025)

Measuring bias
in language models

1. Take a sentence with a target and attribute word
   "He is a kindergarten teacher."

2. Mask the target word
   "[MASK] is a kindergarten teacher."

3. Obtain the probability of target word in the sentence
   $p_T = P(he = [MASK]|sent)$

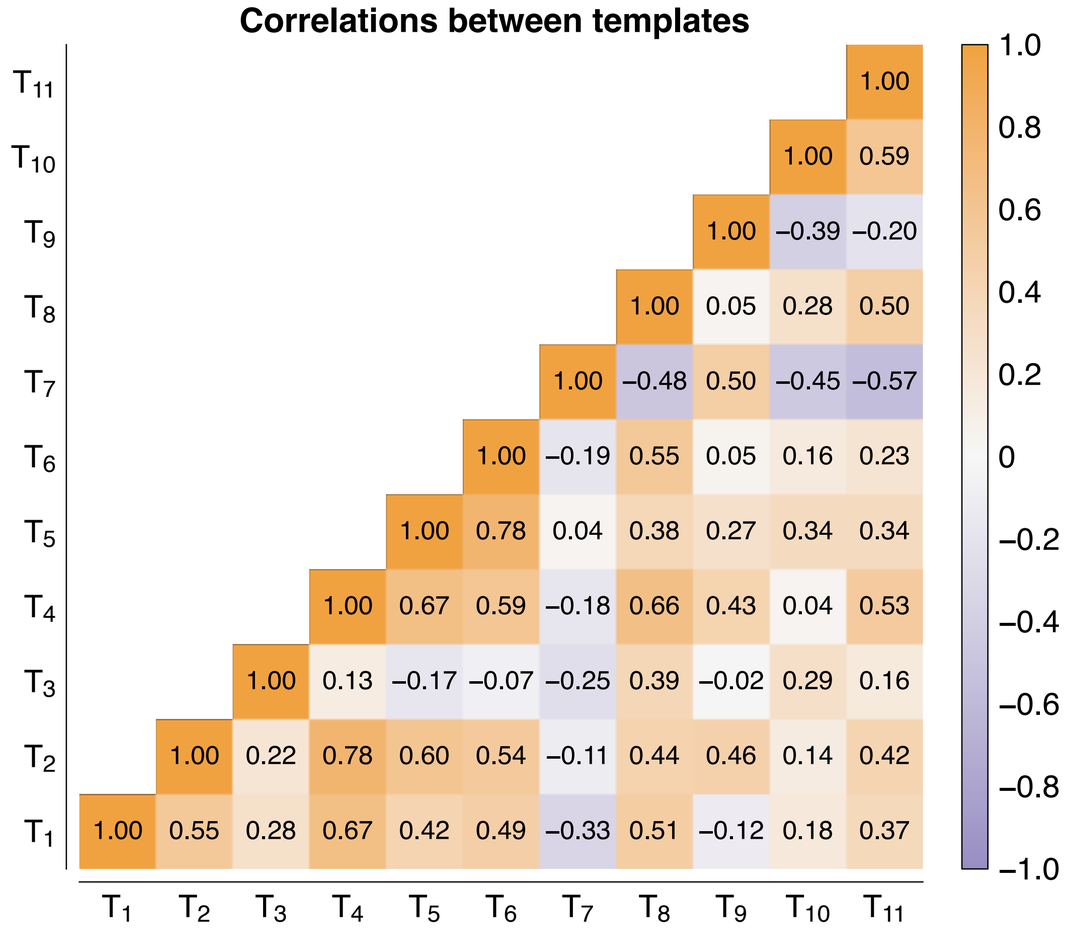
4. Mask both target and attribute word. In compounds, mask each component separately.
   "[MASK] is a [MASK] [MASK]."

5. Obtain the prior probability, i.e. the probability of the target word when the attribute is masked
   $p_{prior} = P(he = [MASK]|masked\_sent)$

6. Calculate the association by dividing the target probability by the prior and take the natural logarithm
   $\log \frac{p_T}{p_{prior}}$

Figure 2: Procedure to calculate the log probability score, after Kurita et al. (2019).

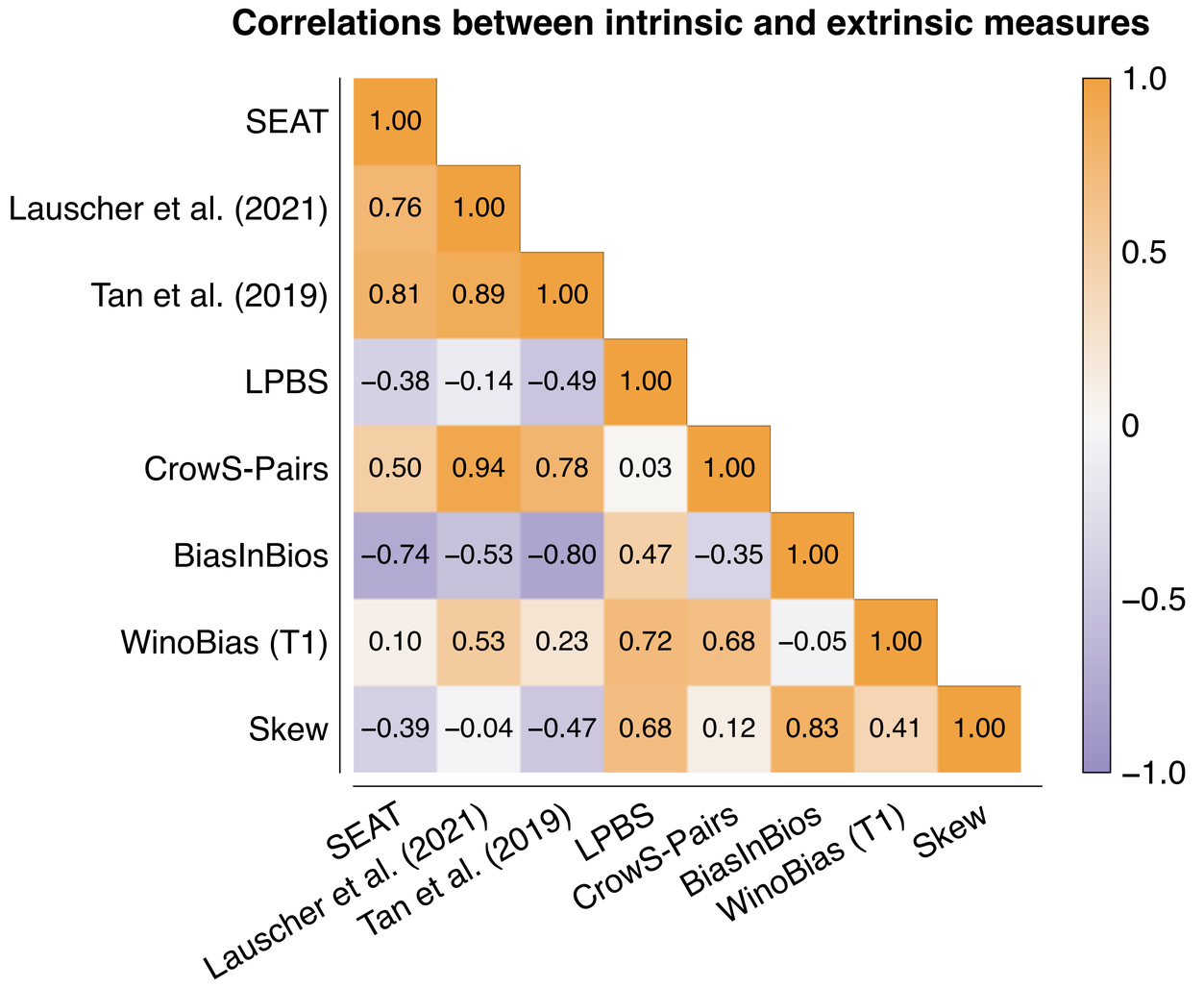
Most templates don't correlate



#	Type	Template sentence	D_{KL}
T_1	Bl.	"This is the _."	—
T_2	Bl.	"That is the _."	0.70
T_3	Bl.	"There is the _."	0.83
T_4	Bl.	"Here is the _."	0.56
T_5	Bl.	"The _ is here."	1.04
T_6	Bl.	"The _ is there."	1.15
T_7	Bl.	"The _ is a person."	2.35
T_8	Bl.	"It is the _."	0.73
T_9	Bl.	"The _ is a `[MASK]`."	2.57
T_{10}	Unbl.	"The _ is an engineer."	4.70
T_{11}	Unbl.	"The _ is a nurse with superior technical skills."	5.02

Delobelle and Tokpo et al. (2022)

… and most metrics don’t correlate



Delobelle and Tokpo et al. (2022)

So what is a ‘good’ metric?

Actionability of metrics

The actual metric does not matter much
SEAT, CEAT, LPBS, DisCo, …

But it needs to test what you care about
e.g. gender bias in professions

Make it explicit what you test

… and test if the metric is reliable
e.g. if different runs yield different results

Metrics for What, Metrics for Whom: Assessing Actionability of Bias Evaluation Metrics in NLP

Pieter Delobelle[1*], Giuseppe Attanasio[2*], Debora Nozza[3], Su Lin Blodgett[4], Zeerak Talat[5]

[1]KU Leuven; Leuven.ai, [2]Instituto de Telecomunicações, Lisbon, [3]MilaNLP, Bocconi
[4]Microsoft Research Montréal, [5]Mohamed bin Zayed University of Artificial Intelligence

Abstract

This paper introduces the concept of *actionability* in the context of bias measures in natural language processing (NLP). We define actionability as the degree to which a measurement's results enable informed action and propose a set of desiderata for assessing it. Building on existing frameworks such as measurement modeling, we argue that actionability is a crucial aspect of bias measures that has been largely overlooked in the literature. We conduct a comprehensive review of 146 papers proposing bias measures in NLP, examining whether and how they provide the information required for actionable results. Our findings reveal that many key elements of actionability, including a measure's intended use and reliability assessment, are often unclear or absent. This study highlights a significant gap in the current approach to developing and reporting bias measures in NLP. We argue that this lack of clarity may impede the effective implementation and utilization of these measures. To address this issue, we offer recommendations for more comprehensive and actionable metric development and reporting practices in NLP bias research.

1 Introduction

As the landscape of bias measures in natural language processing (NLP) has expanded, so too has the literature examining and interrogating these measures (e.g., Blodgett et al., 2021; Goldfarb-Tarrant et al., 2021; Delobelle et al., 2022; Orgad and Belinkov, 2022; Selvam et al., 2023; Goldfarb-

of *validity* and *reliability* for assessing measures (Jacobs and Wallach, 2021; Blodgett et al., 2021).

Across the literature proposing and examining bias measures, talk about measures is often informally tied to talk about what can be done with results produced by measures—i.e., measures' results are often used in decision-making, and good measures should not only exhibit characteristics such as validity and reliability, but should also facilitate decision-making or intervention. For example, natural language generation practitioners use the results of automated metrics to select which models should undergo human evaluation (Zhou et al., 2022b), while other measures' results might guide policies for model release and deployment (Solaiman, 2023). Together, this suggests another piece of vocabulary with which we might assess bias measures. In this paper, we seek to formalize this intuition by introducing *actionability*—the degree to which a measure's results enable informed action—and outlining a set of *desiderata* for actionability—what information is required of a bias measure in order to act based on its results.

At the same time, while the measurement modeling literature has shown the importance of clearly conceptualizing bias and establishing bias measures' validity and reliability, it has also shown that the NLP literature routinely fails to do so. For example, bias in the NLP literature is often underspecified (Blodgett et al., 2020), and measures are often poorly matched to the constructs they are intended to measure (Gonen and Goldberg, 2019; Blodgett

Slides available: pieter.ai/appearances.html