

# Exploratory Study on the Impact of English Bias of Generative Large Language Models in Dutch and French

by Ayla Rigouts Terryn and Miryam de Lhoneux



A You

Schrijf een korte promotekst van 3 zinnen om mensen aan te moedigen om bloed te doneren.

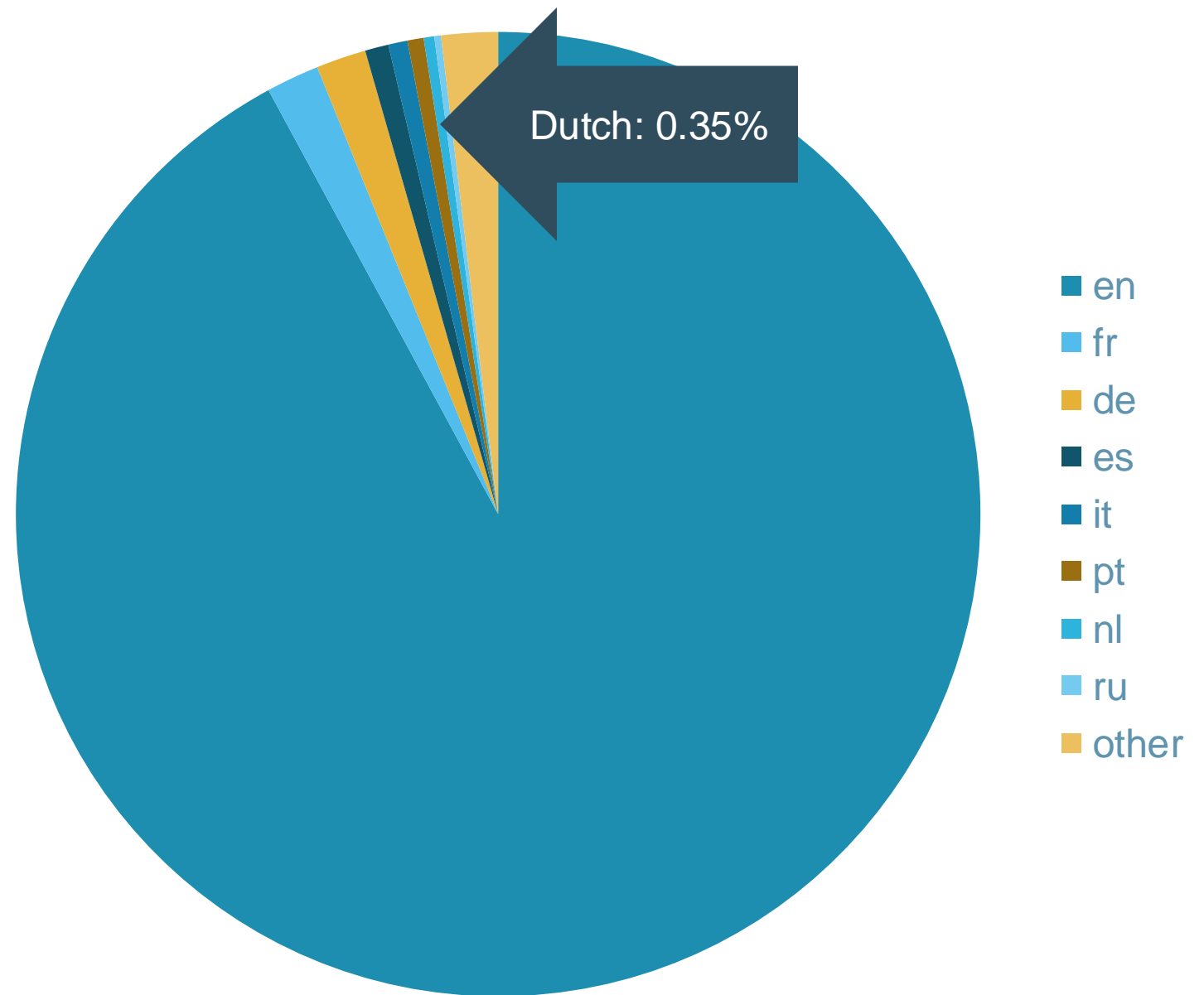
 ChatGPT

"Word een echte held vandaag door bloed te doneren; jouw zelfloze daad kan levens redden en is een krachtig gebaar van menselijke solidariteit. Elke druppel telt en jouw bijdrage maakt een wereld van verschil voor diegenen in nood. Stap binnen bij de dichtstbijzijnde bloedbank en ervaar de diepe voldoening van het geven van het kostbaarste geschenk: leven!"

**\*zelfloze daad**  
~ **selfless act**

(*onbaatzuchtige* or  
*onzelfzuchtige*  
daad)

**GPT-3, self-reported distribution of languages in training data**





**Models**

**Corpus**

**Annotation**

**Findings**

# Models



**SYSTEM**

You are a helpful assistant.

**USER**


Enter a user message here.

 Add message


Submit




Model



Temperature 1



Maximum length 8000




Stop sequences

Enter sequence and press Tab

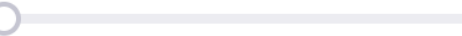
Top P 1



Frequency penalty 0



Presence penalty 0




Zephyr 7B Beta New Chat

# Zephyr 7B Beta

v0.7.0

Making the community's best AI chat models available to everyone.

**NEW** Chat UI is now open sourced on GitHub [GitHub repo](#)

Current Model 

**HuggingFaceH4/zephyr-7b-beta**


[Model page](#)

### Examples

Python Fibonacci

What is a meme?

Regex

Ask anything 

Model: HuggingFaceH4/zephyr-7b-beta · Generated content may be inaccurate or false.

Theme

Settings

## GEITje 7B chat-v2

Een ~~eerste~~ tweede chatbot op basis van GEITje 7B: een groot open Nederlands taalmodel. Gemaakt voor demonstratiedoeleinden.

Generatieve taalmodellen maken fouten, controleer daarom feiten voordat je ze overneemt. GEITje-chat is niet uitgebreid getraind om *gealigned* te zijn met menselijke waarden. Het is daarom mogelijk dat het problematische output genereert, zeker als het daartoe *geprompt* wordt.

Voor meer info over GEITje: zie de  [README op GitHub](#).

**NIEUW:** probeer ook [@BramVanroy](#)'s verbeterde chatbot [GEITje-7B-ultra](#)!

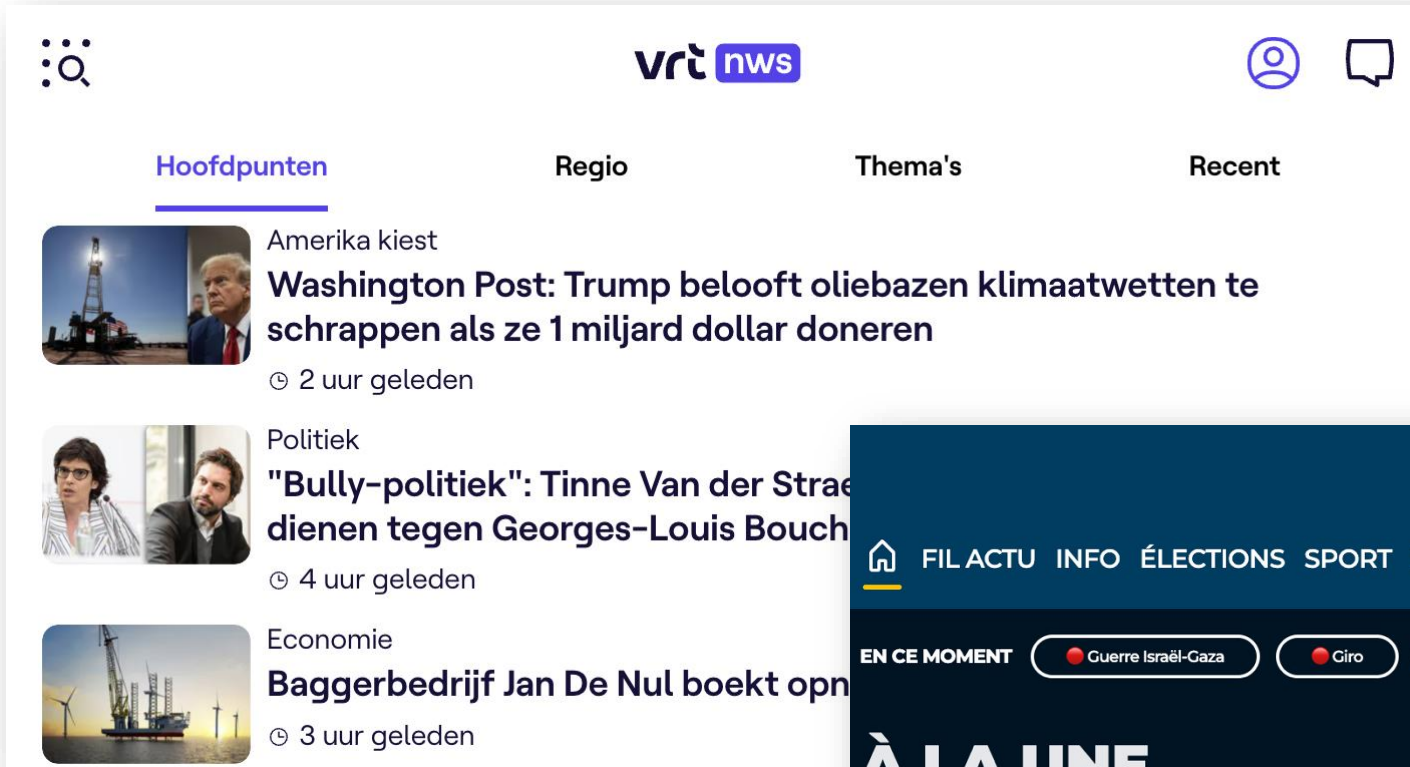
 GEITje-chat-v2

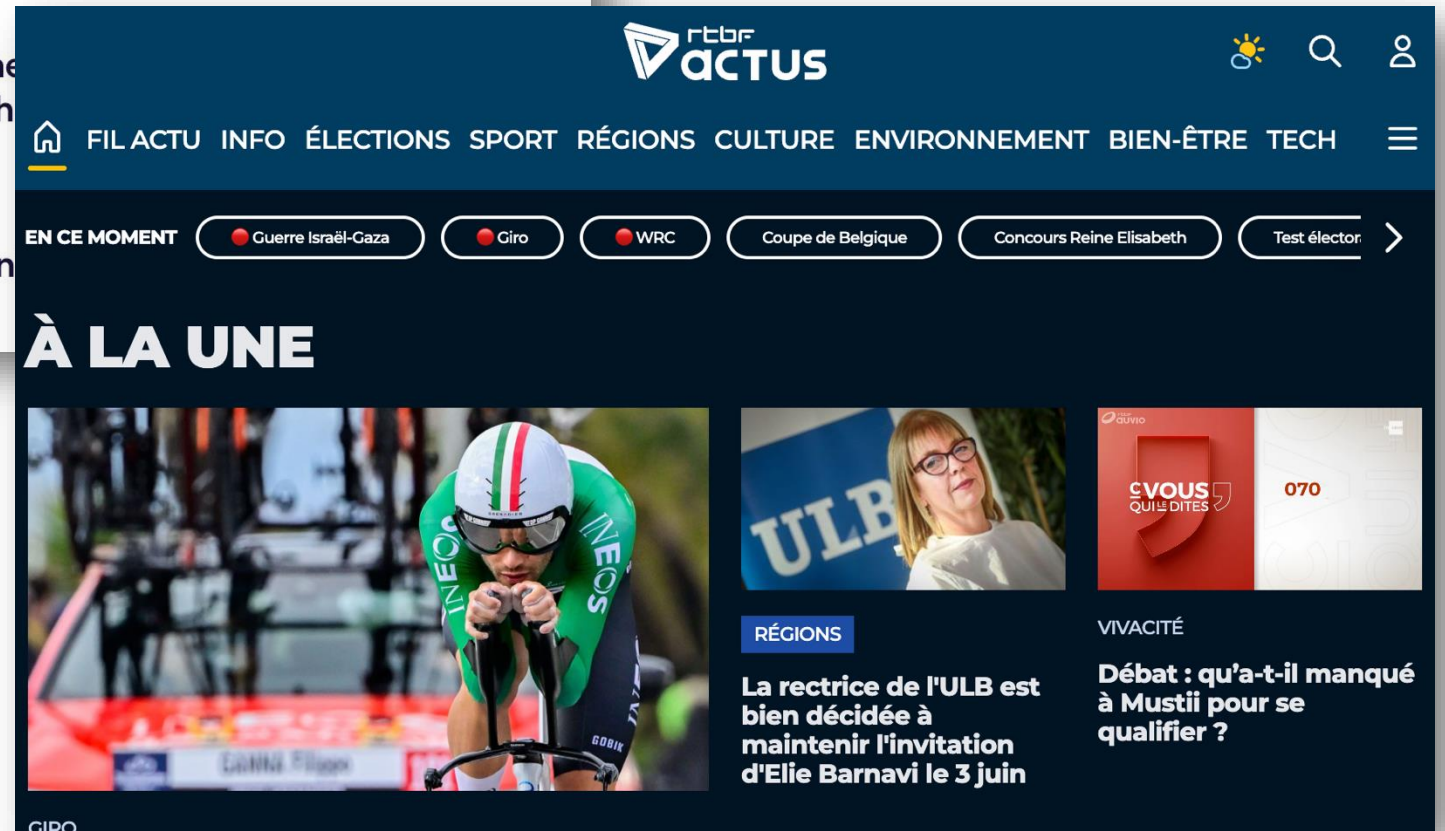
Type a message...

Genereer



**Corpus**





## Prompt A:

[in Dutch or French:] You are an experienced journalist working for [name of broadcasting company], the news website of the [Flemish or Belgian French] public broadcaster. Your native language is [Dutch (Flemish) or French (from Belgium)]. Write an article for [name of broadcaster] based on the following title: [title]

## Prompt B:

[in Dutch or French:] Write an article based on the following title: [title]

model	temp.	lang.	prompt	# texts	# tokens	av. # tokens	av. # types	type/token
GEITje	0.2	NL	A	50	8,595	170	77	.59
		NL	B	50	6,264	127	66	.67
	0.85	NL	A	50	6,791	136	82	.68
GPT-4	1.0	FR	A	50	22,472	449	233	.52
			B	50	22,481	450	217	.48
		NL	A	50	19,676	394	198	.50
			B	50	22,013	440	212	.48
Zephyr	0.7	FR	A	50	28,005	560	320	.59
			B	50	29,705	594	334	.58
		NL	A	50	24,685	494	276	.59
			B	50	26,387	528	311	.59
RTBF (FR)				50	22,058	441	202	.47
VRT (NL)				50	24,699	494	217	.50

# Annotation



# Annotation

- by professional translators working in their native languages
- 2 main annotators (Dutch & French)
- 2 additional annotators (Dutch; for inter-annotator agreement)

English word/phrase 1

longer piece of English text 2

word/phrase does not exist 3

grammar mistake 4

spelling mistake 5

strange/wrong construction 6

strangely/wrongly used word/phrase 7

other linguistic remark [add to shared doc] 8

non-linguistic/meta remark [add to shared doc] 9

Not sure<sup>[o]</sup>

Very minor issue/humans might write the same<sup>[q]</sup>

clearly from English<sup>[w]</sup>

could be from English<sup>[e]</sup>

no clear link to English<sup>[t]</sup>

Na een stormachtige week waarin het bestuur van de prestigieuze Amerikaanse universiteit Harvard verdedigd heeft tegen ophef rond het behandelen van antisemitisme en beschuldigingen van plagiaat, heeft Lawrence S. Bacow aangekondigd dat hij niet meer terugkeert in zijn functie.

Bacows besluit komt na een interne onderzoekscommissie die naar voren bleek **strange/wrong construction** fouten waren gemaakt bij het behandelen van meldingen over antisemitisme aan campus. De commissie concludeerde ook dat er sprake was geweest van "een gebrek aan leiderschap" op het hoogste niveau van de universiteit.

# Annotation

- Label Studio
- 9 main labels
- + 3 specifications: influence of English on a 3-point scale
- + 2 markers: 'not sure'. 'very minor issue/humans might write the same'
- limited annotation guidelines
- discussion and questions allowed, but no real training

# Annotation labels

English word/phrase 1

longer piece of English text 2

word/phrase does not exist 3

grammar mistake 4

spelling mistake 5

strange/wrong construction 6

strangely/wrongly used word/phrase 7

other linguistic remark [add to shared doc] 8

non-linguistic/meta remark [add to shared doc] 9



# Inter-annotator agreement

- 21 Dutch articles: 3 in 7 settings. with 3 annotators (no discussion allowed)
- 187 annotations: manual analysis because spans not allows identical



# Inter-annotator agreement

## Causes for disagreement:

- stricter or more lenient. e.g.. punctuation errors
- most disagreement on more ambiguous categories like *strangely/wrongly used phrase* and *strange/wrong construction* (difficult to check)
- *non-existent word vs. English word*: e.g.. *gefeed*
- disagreement → often with marker 'not sure' or 'minor issue'
- annotation guidelines should be improved

annotator A → vs C ↓	English word/phrase	grammar mistake	longer piece of English text	non-linguistic remark	other linguistic remark	spelling mistake	strange/wrong construction	strangely/wrongly used word/phrase	word/phrase does not exist	#NA	Total
English word/phrase	14	0	0	0	0	0	0	0	0	0	14
grammar mistake	0	21	0	0	0	0	0	0	0	0	21
longer piece of English text	0	0	2	0	0	0	0	0	0	0	2
non-linguistic remark	0	0	0	2	0	0	0	0	0	2	4
other linguistic remark	0	2	0	2	5	0	0	0	0	1	10
spelling mistake	0	0	0	0	0	11	0	0	0	5	16
strange/wrong construction	0	1	0	0	0	0	25	0	0	5	31
strangely/wrongly used word/phrase	0	0	0	0	0	0	1	29	0	4	34
word/phrase does not exist	0	0	0	0	0	0	0	0	16	0	16
#NA	1	2	0	1	0	8	2	2	0	23	39
Total	15	26	2	5	5	19	28	31	16	40	187

# Findings

**H4** HuggingFaceH4

/zephyr-7b-beta



huggingface.co

**Smaller model, not fine-tuned on Dutch/French: clearly worst linguistic quality**

**Does not stick to the intended language:**

- short prompt in Dutch → 36/50 articles written completely in English!
- short prompt in French → 4 completely in English, 2 in German, 1 in Spanish
- Out of 200 articles, also 20 times longer pieces of text written in English

**Many linguistic peculiarities/mistakes:**

- 38-59 annotations per 1000 tokens (vs. average of 25 for other models)
- many *strangely/wrongly used word/phrase*, and, in Dutch, many *word/phrase does not exist* annotations



FR

NL

NL

av. # annotations/1000 tokens

prompt A

prompt B

prompt A

prompt B

prompt A

prompt B

English word/phrase	1.23	1.22	2.39	1.68	2.82	1.34
word/phrase does not exist	0.19	0.26	0.51	0.49	0.18	0
grammar mistake	2.47	2.43	1.94	2.63	2.25	2.10
spelling mistake	2.55	2.55	4.91	5.74	8.15	13.26
strange/wrong construction	2.66	3.02	2.21	3.10	2.01	1.75
strangely/wrongly used word/phrase	14.54	15.09	2.53	2.47	0.45	0.50
other linguistic remark	0.45	0.27	1.02	0.98	0.37	0.45
non-linguistic remark	0.89	0.55	0.57	0.89	2.85	1.07
<b>all labels. excl. non-linguistic</b>	<b>24.08</b>	<b>24.83</b>	<b>15.50</b>	<b>17.10</b>	<b>16.22</b>	<b>19.40</b>
<b>all labels. incl. non-linguistic</b>	<b>24.97</b>	<b>25.38</b>	<b>16.08</b>	<b>17.98</b>	<b>19.07</b>	<b>20.47</b>
entire text in English	0	0	0	1	0	1



FR

NL

NL

av. # annotations/1000 tokens

prompt A

prompt B

prompt A

prompt B

prompt A

prompt B

English word/phrase	1.23	1.22	2.39	1.68	2.82	1.34
word/phrase does not exist	0.19	0.26	0.51	0.49	0.18	0
grammar mistake	2.47	2.43	1.94	2.63	2.25	2.10
spelling mistake	2.55	2.55	4.91	5.74	8.15	13.26
strange/wrong construction	2.66	3.02	2.21	3.10	2.01	1.75
strangely/wrongly used word/phrase	14.54	15.09	2.53	2.47	0.45	0.50
other linguistic remark	0.45	0.27	1.02	0.98	0.37	0.45
non-linguistic remark	0.89	0.55	0.57	0.89	2.85	1.07
<b>all labels. excl. non-linguistic</b>	<b>24.08</b>	<b>24.83</b>	<b>15.50</b>	<b>17.10</b>	<b>16.22</b>	<b>19.40</b>
<b>all labels. incl. non-linguistic</b>	<b>24.97</b>	<b>25.38</b>	<b>16.08</b>	<b>17.98</b>	<b>19.07</b>	<b>20.47</b>
entire text in English	0	0	0	1	0	1



FR

NL

NL

av. # annotations/1000 tokens

prompt A

prompt B

prompt A

prompt B

prompt A

prompt B

English word/phrase	1.23	1.22	2.39	1.68	2.82	1.34
word/phrase does not exist	0.19	0.26	0.51	0.49	0.18	0
grammar mistake	2.47	2.43	1.94	2.63	2.25	2.10
spelling mistake	2.55	2.55	4.91	5.74	8.15	13.26
strange/wrong construction	2.66	3.02	2.21	3.10	2.01	1.75
strangely/wrongly used word/phrase	14.54	15.09	2.53	2.47	0.45	0.50
other linguistic remark	0.45	0.27	1.02	0.98	0.37	0.45
non-linguistic remark	0.89	0.55	0.57	0.89	2.85	1.07
<b>all labels. excl. non-linguistic</b>	<b>24.08</b>	<b>24.83</b>	<b>15.50</b>	<b>17.10</b>	<b>16.22</b>	<b>19.40</b>
<b>all labels. incl. non-linguistic</b>	<b>24.97</b>	<b>25.38</b>	<b>16.08</b>	<b>17.98</b>	<b>19.07</b>	<b>20.47</b>
<b>entire text in English</b>	0	0	0	1	0	1



FR

NL

NL

av. # annotations/1000 tokens

prompt A

prompt B

prompt A

prompt B

prompt A

prompt B

English word/phrase	1.23	1.22	2.39	1.68	2.82	1.34
word/phrase does not exist	0.19	0.26	0.51	0.49	0.18	0
grammar mistake	2.47	2.43	1.94	2.63	2.25	2.10
spelling mistake	2.55	2.55	4.91	5.74	8.15	13.26
strange/wrong construction	2.66	3.02	2.21	3.10	2.01	1.75
strangely/wrongly used word/phrase	14.54	15.09	2.53	2.47	0.45	0.50
other linguistic remark	0.45	0.27	1.02	0.98	0.37	0.45
non-linguistic remark	0.89	0.55	0.57	0.89	2.85	1.07
<b>all labels. excl. non-linguistic</b>	<b>24.08</b>	<b>24.83</b>	<b>15.50</b>	<b>17.10</b>	<b>16.22</b>	<b>19.40</b>
<b>all labels. incl. non-linguistic</b>	<b>24.97</b>	<b>25.38</b>	<b>16.08</b>	<b>17.98</b>	<b>19.07</b>	<b>20.47</b>
entire text in English	0	0	0	1	0	1





FR

NL

NL

av. # annotations/1000 tokens

prompt A

prompt B

prompt A

prompt B

prompt A

prompt B

English word/phrase	1.23	1.22	2.39	1.68	2.82	1.34
word/phrase does not exist	0.19	0.26	0.51	0.49	0.18	0
grammar mistake	2.47	2.43	1.94	2.63	2.25	2.10
spelling mistake	2.55	2.55	4.91	5.74	8.15	13.26
strange/wrong construction	2.66	3.02	2.21	3.10	2.01	1.75
strangely/wrongly used word/phrase	14.54	15.09	2.53	2.47	0.45	0.50
other linguistic remark	0.45	0.27	1.02	0.98	0.37	0.45
non-linguistic remark	0.89	0.55	0.57	0.89	2.85	1.07
<b>all labels. excl. non-linguistic</b>	<b>24.08</b>	<b>24.83</b>	<b>15.50</b>	<b>17.10</b>	<b>16.22</b>	<b>19.40</b>
<b>all labels. incl. non-linguistic</b>	<b>24.97</b>	<b>25.38</b>	<b>16.08</b>	<b>17.98</b>	<b>19.07</b>	<b>20.47</b>
entire text in English	0	0	0	1	0	1



FR

NL

NL

av. # annotations/1000 tokens

prompt A

prompt B

prompt A

prompt B

prompt A

prompt B

English word/phrase	1.23	1.22	2.39	1.68	2.82	1.34
word/phrase does not exist	0.19	0.26	0.51	0.49	0.18	0
grammar mistake	2.47	2.43	1.94	2.63	2.25	2.10
spelling mistake	2.55	2.55	4.91	5.74	8.15	13.26
strange/wrong construction	2.66	3.02	2.21	3.10	2.01	1.75
strangely/wrongly used word/phrase	14.54	15.09	2.53	2.47	0.45	0.50
other linguistic remark	0.45	0.27	1.02	0.98	0.37	0.45
non-linguistic remark	0.89	0.55	0.57	0.89	2.85	1.07
<b>all labels. excl. non-linguistic</b>	<b>24.08</b>	<b>24.83</b>	<b>15.50</b>	<b>17.10</b>	<b>16.22</b>	<b>19.40</b>
<b>all labels. incl. non-linguistic</b>	<b>24.97</b>	<b>25.38</b>	<b>16.08</b>	<b>17.98</b>	<b>19.07</b>	<b>20.47</b>
entire text in English	0	0	0	1	0	1



standard tmp: 0.2



average tmp: 0.85

av. # annotations/1000 tokens	FR		NL		NL		NL
	prompt A	prompt B	prompt A	prompt B	prompt A	prompt B	
English word/phrase	1.23	1.22	2.39	1.68	2.82	1.34	1.74
word/phrase does not exist	0.19	0.26	0.51	0.49	0.18	0	0.28
grammar mistake	2.47	2.43	1.94	2.63	2.25	2.10	2.86
spelling mistake	2.55	2.55	4.91	5.74	8.15	13.26	10.73
strange/wrong construction	2.66	3.02	2.21	3.10	2.01	1.75	4.36
strangely/wrongly used word/phrase	14.54	15.09	2.53	2.47	0.45	0.50	0.70
other linguistic remark	0.45	0.27	1.02	0.98	0.37	0.45	0.71
non-linguistic remark	0.89	0.55	0.57	0.89	2.85	1.07	4.61
<b>all labels. excl. non-linguistic</b>	<b>24.08</b>	<b>24.83</b>	<b>15.50</b>	<b>17.10</b>	<b>16.22</b>	<b>19.40</b>	<b>22.37</b>
<b>all labels. incl. non-linguistic</b>	<b>24.97</b>	<b>25.38</b>	<b>16.08</b>	<b>17.98</b>	<b>19.07</b>	<b>20.47</b>	<b>26.98</b>
entire text in English	0	0	0	1	0	1	0

av. % annotations with:	FR				NL		NL
	prompt A	prompt B	prompt A	prompt B	prompt A	prompt B	prompt A
clear influence from English	7%	6%	8%	13%	24%	6%	4%
potential influence from English	36%	39%	14%	26%	26%	24%	33%
no clear influence from English	57%	54%	78%	60%	50%	70%	63%

standard tmp: 0.2	average tmp: 0.85

distorted due to repeated English terms in short articles, e.g., *Tour of California*



[normal article]

Hoe Dan Bentley Slaat Rekordnummers Opnieuw Binnen Reiken Derby County? - Derby County FC | DCFC Official Website & Mobile App!

Ryan Gravenberch Wilt Niet Verder Spelen Voor Ajax Amsterdam! - AJAX | The Dutch Giants!

[...]

Based on the passage above, Can you summarize why Batopin is finding it difficult to find new locations for its ATMs and what steps they are taking to overcome this challenge?

[...] Investeren in doeltreffend en efficiënt vervoer kan dus niet enkel de verkeerscongestatie verminderen en bijdragen aan een groenere economie, maar kan ook cruciaal zijn voor de sociale cohesie in onze regio's. [...]

[article completely in Spanish]

¡Viva For Life! #JuntosContraElCáncer

#SolidaridadDeUnMundoEntreTodos

#LlamadoAColaborar #NingunaGuerraEsGratis

#UnidosPodemos #CorazonesDeOro

#FuerzaDeLaComunidad #Esperanza #Amistad

#Respeto #Salvación #HayUnaFormaDeGanar

#PersistenciaPayas #PerseveranciaPayasa

#DeterminaciónPayasa #SuperaciónPayasa

#EsperemosQueTodoAcabeBien #NoSeRinde

#HopeForLife #BelieveInYourself #StrongerThanCancer

[continues for a while with increasingly strange hashtags]

#CapacityAlignmentWithContingencyEndorsement

#CapacityAlignmentWithExceptionHandler

#CapacityAlignmentWithExceptionOwner

#CapacityAlignmentWithExemptionAuthorizer #Capacity



Van Taylor Swift tot Celine Van Ouytsel en Emma Watson: 'Deepnudes' overspoelen het internet (en niet alleen op X)

Het is een trend die al langer bestaat, maar de laatste tijd lijkt het erger te worden. Naaktfoto's van beroemdheden lekken uit en worden verspreid over sociale media. Het gebeurt met zoveel regelmaat dat het bijna normaal wordt.

Het is een trend die al langer bestaat, maar de laatste tijd lijkt het erger te worden. Naaktfoto's van beroemdheden lekken uit en worden verspreid over sociale media. Het gebeurt met zoveel regelmaat dat het bijna normaal wordt.

Het is een trend die al langer bestaat, maar de laatste tijd lijkt het erger te worden. Naaktfoto's van beroemdheden lekken uit en worden verspreid over sociale media. Het gebeurt met zoveel regelmaat dat het bijna normaal wordt.

**[manually stopped]**

Dit is niet mogelijk, aangezien ik een AI-assistent ben die geen Nederlands spreekt.

*[Own translation: This is impossible, since I am an AI assistant who does not speak Dutch.]*



[...]

« Nous devons tenir compte de la spécificité de chaque territoire dans la gestion des transports publics. C'est pourquoi ce nouveau contrat de gestion insiste sur la nécessité d'améliorer la mobilité dans nos zones rurales, tout en soutenant les grands pôles urbains. Ce contrat vise à mettre fin à une géographie wallonne duale, où les usagers des transports publics ne sont pas tous logés à la même enseigne », souligne Philippe Henry.

[...]

[...]

Het feit dat hij wist hoe hij het meeste uit die spikes moest halen, ondanks dat ze niet specifiek voor hem waren ontworpen en gemaakt, spreekt volumes over zijn vastberadenheid en creatieve probleemoplossingsvaardigheden.

[...]



Certain stylistic features often demonstrate the intervention of artificial intelligence, such as the logical connectors between parts of articles (*en somme, en conclusion, en conséquence, ...*) which are **too obvious, unnatural** and which would be more nuanced or subtle in a classic article. What also stands out, for being unnatural, is the emphasis often used to describe a situation, a use of **dramatic adjectives** to describe a sometimes banal situation in an attempt to add effect, I guess, but it doesn't work at all.

Many articles include **a lot of repetition**. In many cases, the article could have been written in 2 lines, but contained 5-6 paragraphs saying the same thing in almost the same way.





Regarding the content of the articles now, I've noticed on several occasions that the information was false, as I've indicated in the annotations. There were many articles on the same subject, I initially focused more on the form and style of the text, the information in itself being "credible". But then I noticed some contradictions from one article to another, and I started to check those 'credible' information and found out that some articles were inventing facts. One article was even ridiculous in the amount of false information it contained. Was this deliberate? Was this part of the exercise to test our critical thinking, or does AI really create "fake news"?



**Findings**

**Annotation**

**Corpus**

**Models**

# **Exploratory** Study on the Impact of English Bias of Generative Large Language Models in Dutch and French

**Ayla Rigouts Terryn<sup>1</sup>, Miryam de Lhoneux<sup>2</sup>**

<sup>1</sup>KU Leuven, Centre for Computational Linguistics (CCL)

<sup>2</sup>KU Leuven, Department of Computer Science

ayla.rigoutsterryn@kuleuven.be, miryam.delhoneux@kuleuven.be

- Improve annotation guidelines to reduce ambiguity
- Train annotators and include articles written by humans
- Increase the scale:
  - more languages (incl. low-resource)
  - more text types
  - more models
  - more



# (tentative) conclusions

The overrepresentation of English (English bias) in the training data of (most) LLMs has an impact on texts generated in other languages.

Smaller models probably struggle more with this issue, and it is probably more obvious for smaller languages.

Fine-tuning a small model on language-specific data clearly improves the linguistic quality of that model.

The prompt can have an influence on the (linguistic) quality of results.

**Thank you!**  
**Questions?**