Schedule

10:00	
	Introduction
10:30	Introduction
10:45	
11:30	Intrinsic bias
	Practical ses
12:00	Lunch
13:30	
13-45	Finetuning
	Extrinsic bias
14:45	Dractical sos
15:30	r actical Ses
46.00	Final though
10:00	



to LLMs Pieter to fairness Ewoenam Pieter ssion Ewoenam Pieter Ewoenam S Pieter ssion nts: legal, ... Pieter

1

Fairness and bias in NLP



- VAIA lecture June 13, 2024
 - Intro







Tell me a fun fact about the Roman Empire

Show me a code snippet of a website's sticky header

?

ChatGPT in the news

Wetenschap

Bewegingsmethode voor kleuters laat lijst 'wetenschappelijke' artikels produceren door ChatGPT



Biba & Loeba © (c) - VRT - Biba & Loeba



3 The Brussels Times ChatGPT diagnoses emergency room patients as well as a doctor, study finds



Waarom ChatGPT vaak empathischer is dan uw dokter'

Ann Peuteman

19-09-2023, 05:00 .

'Patiënten willen in de eerste plaats erkenning, en die krijgen ze vandaag blijkbaar eerder van een chatbot dan van een echte arts', schrijft Knack-redactrice Ann Peuteman in haar column De Zoetzure Dinsdag.





How does this work?

How does this work? How can I use this for my research?





Parts of a language models

'Heads' of a language model

How a model predicts the next word

Attention mechanism

Each word affects the other words

Tokenizer

How a model understands text

Training data

What a model learns





Training data



(copyright free) books

wikipedia







No, I am not a giraffe.



No, I am not a giraffe.

No, I am not a giraffe.

No, I am not a giraffe.

No, I am not a giraffe. [2822, 11, 358, 1097, 539, 264, 41389, 38880, 13]

No, I am not a giraffe.

No, I am not a giraffe. [2822, 11, 358, 1097, 539, 264, 41389, 38880, 13]



21223: affe





Embeddings capture meaning

0.9 0.1 0.1 0.5 0.4 0.1 0.0

Giraffe





Horse

Embeddings capture meaning Word embeddings

Norway

. . .



Horse	Word	Cosine	distance
	norway denmark finland		0.760124 0.715460 0.620022
	switzerland belgium netherlands		0.588132 0.585839 0.574631
	estonia slovenia		0.562368 0.547621 0.531408



















Bank















Bank

















Bank

















Bank







Predicting the next token

It is the tallest living terrestrial animal.

Giraffes live in herds.

He is a giraffe. IUCN recognises one species of giraffe.









Large training corpuses are used





One book 40-50k tokens



One LLM training set

2.5T - 6T tokens ~2 500 000 bookshelves

One bookshelf

1.6M - 2.5M tokens



Pretraining is expensive, but worth it



https://openai.com/research/scaling-laws-for-neural-language-models https://allenai.org/olmo





Language modeling



1. Causal language modeling (CLM)





2. Masked language modeling (MLM) He is a doctor Masked LM He <m> a doctor



Language modeling



1. Causal language modeling (CLM)





RobBERT



He is a doctor Masked LM *He <m> a doctor*

https://pieter.ai/robbert/

2. Masked language modeling (MLM)



Fairness and bias in NLP

VAIA lecture June 13, 2024

Intrinsic bias



What is the problem?

🔁 Fill-Mask

Mask token: [MASK]

[MASK] is a nurse.

Compute

Computation time on Intel Xeon 3rd Gen Scalable cpu: 0.038 s

she	0.867
he	0.013
kim	0.001
• sarah	0.001
maria	0.001

🔁 Fill-Mask

Mask token: [MASK]

[MASK] is a professor.

Compute

Computation time on Intel Xeon 3rd Gen Scalable cpu: 0.040 s

he	Θ.ε	838
she	0.1	129
it	Θ.0	902
• his	0.0	000
and	Θ.0	900







Measuring bias in non-contextual word embeddings



Target words

KU LEUVEN



Measuring bias in non-contextual word embeddings



KU LEUVEN





Measuring bias in non-contextual word embeddings



KU LEUVEN





1. Pretraining step e.g. OSCAR, Wikipedia, ...







1. Pretraining step e.g. OSCAR, Wikipedia, ...







1. Pretraining step e.g. OSCAR, Wikipedia, ...





2. Finetuning step e.g. sentiment analysis, named entity recognition Transfer [CLS] learning **Finetuned model** e.g. BERT He is a doctor



1. Pretraining step e.g. OSCAR, Wikipedia, ...





2. Finetuning step e.g. sentiment analysis, named entity recognition Transfer [CLS] learning **Finetuned model** e.g. BERT He is a doctor **Extrinsic biases**


Measuring bias in MLMs

1. Pretraining step e.g. OSCAR, Wikipedia, ...





2. Finetuning step e.g. sentiment analysis, named entity recognition Transfer [CLS] learning **Finetuned model** e.g. BERT He is a doctor **Extrinsic biases**

https://fairlearn.org/



Intrinsic bias in MLMs

Measuring fairness in LMs

Background on biases

Intrinsic biases in language models

1. Pretraining step e.g. OSCAR, Wikipedia, ...





Mitigating stereotypes in LMs

Conclusion

2. Finetuning step e.g. sentiment analysis, named entity recognition





Nikita Nangia*	Clara	Vania*	Rasika Bhalerao	* Samuel R. Bowman
likita i tangin		New Y	ork University	, Januar odu
{nikitanan	gia,	c.vania	, rasikabh,	bowman}@nyu.euu

Abstract

Warning: This paper contains explicit statements of offensive stereotypes and may be upsetting.

Pretrained language models, especially masked language models (MLMs) have seen success across many NLP tasks. However, there is ample evidence that they use the cultural biases that are undoubtedly present learn and use these biases (Bolukbasi et al., 2016; Caliskan et al., 2017; Garg et al., 2017; May et al., 2010; Zhao et al., 2018; Rudinger et al., 2017). Models that have learnt representations that are biased against historically disadvantaged groups can cause a great deal of harm when those biases surface in downstream tasks or applications, such as automatic summarization or web search (Bender, 2019). Identifying and quantifying the learnt biases enables us to measure progress as we build less bipeople, e.g., Asians ns are bad drivers. known to hurt tarhed language modal world data, they otypical biases. In ects of these modtify the bias caperature on quantied language modcially constructed We present Steretaset in English to in four domains: nd religion. We e BERT, GPT2, our dataset and bit strong stereoent a leaderboard the bias of future

Reducing Gendered Correlations in Pre-t	rained Models
suring Bias in Contextualized Word Rep	resentations
Curita Nidhi Vyos Ayush Beresh Alan W Di L	
Carnegie Mellon University ta,nkvyas,apareek,awb,ytsvetko}@an	Idrew.cmu.e On Measuring Social Biases in Se
Abstract embeddings such as REPT	Chandler May ¹ Alex Wang ² S
bias in pretrained language models Bethke [†] and Siva Reddy [‡] chnology, Cambridge MA, USA	Assessing Social and Intersectional Biases in Contextualized Word Representations
Gill University, Montreal, QC, Canada na.bethke@intel.com, mila.quebec	Yi Chern Tan, L. Elisa Celis Yale University {yichern.tan, elisa.celis}@yale.edu
Choose the appropriate word: Domain: Gender Target: Girl Context: Girls tend to be more than boys	Abstract Social bias in machine learning has drawn significant attention, with work ranging from demonstrations of bias in a multitude of applications, curating definitions of fairness for different contexts, to developing algorithms to mitigate bias. In natural language processing, gender bias has been shown to exist in context-free word embeddings. Recently, contextual word representations have outperformed word embeddings in several downstream NLP tasks. These word representations the entire sentence. In this paper, we analyze the extent to which state-of-the-art models for contextual word representations, such as BERT and GPT-2, encode biases with respect to gender, race, and intersectional identities. Towards this, we propose assessing bias at the contextual word level. This novel approach captures the contextual effects of bias missing in context-free word embeddings, yet avoids contextual effects that underestimate bias at the sentence encoding level. We demonstrate evidence of bias at the corpus level, find varying evidence of bias in ontextual word models, and observe that bias effects for intersectional minorities are exacerbated beyond their constituent minority identities. Further, evaluating is effects at the contextual word level captures biases that are not captured at the sentence level, confirming the need for our novel approach.
(b) The Intersentence Context Association Test	Introduction
Figure 1: Context Association Tests (CATs) to mea- sure the bias and language modeling ability of language	d embeddings [22, 24], which provide context-free vector representations of words, have become dard practice in NLP. Recently, contextual word representations [19, 17, 25, 26, 10, 27] have had

entence Encoder Shikha Bordia² inger¹ University 16, bowman to phrases and Association Te ences generated m Caliskan et h as "This is a[strate the new p h and advance lso introduce t enable to word ack woman ster 09; Harris-Pen (16) and a dou settings (Heili of sentence-lev the impact of or example, sev n given name ican and Afric rms referring ch as "woman" of using given 1 g alternate ver the two. This T, as categori non single-wo a varying evi

Measuring fairness in LMs Mitigating stereotypes in LMs Conclusion Measuring bias in word embeddings



Bias subspaces (based on PCA) Bolukbasi et al. (2016)



Targets (N)	Attributes (N)	GIONe*	word2vec
Flowers vs Insects (25)	Pleasant vs Unpleasant (25)	1.50*	1.54*
Instruments vs Weapons (25)	Pleasant vs Unpleasant (25)	1.53*	1.63*

Word embeddings association tests Caliskan et al. (2017)



Word Embeddings Association Test (WEAT)

- targets X,Y
 - Targets: e.g. man, woman
 - Attributes: e.g. doctor, nurse



Conclusion

Measure mean cosine distance between attributes A,B and

$$s(\mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{B}) = \sum_{x \in \mathcal{X}} u(x, \mathcal{A}, \mathcal{B}) - \sum_{y \in \mathcal{Y}} u(y, \mathcal{A}, \mathcal{A})$$

Caliskan et al. (2017)





Word Embeddings Association Test (WEAT)

- targets X,Y
 - Targets: e.g. man, womar
 - Attributes: e.g. doctor, nurse
- Based on the *implicit* association test



Conclusion

• Measure mean cosine distance between attributes A,B and

$$s(\mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{B}) = \sum_{x \in \mathcal{X}} u(x, \mathcal{A}, \mathcal{B}) - \sum_{y \in \mathcal{Y}} u(y, \mathcal{A}, \mathcal{B})$$

Caliskan et al. (2017)





How about language models?

Measuring fairness in LMs

Many choices that result in different metrics

Step 1 Choose your base fairness metric



Targets (N)	Attributes (N)	GION ^{e*}	word?wec
Flowers vs Insects (25)	Pleasant vs Unpleasant (25)	1.50*	1.54*
Instruments vs Weapons (25)	Pleasant vs Unpleasant (25)	1.53*	1.63*

PCA Bolukbasi et al. (2016)

WEAT Caliskan et al. (2017)





Conclusion

Measuring fairness in LMs

Many choices that result in different metrics

Step 1 Choose your base fairness metric



Targets (N)	Attributes (N)	GION ^{e*}	word?wec
Flowers vs Insects (25)	Pleasant vs Unpleasant (25)	1.50*	1.54*
Instruments vs Weapons (25)	Pleasant vs Unpleasant (25)	1.53*	1.63*

PCA Bolukbasi et al. (2016)

WEAT Caliskan et al. (2017)



Step 2 Select all the templates you want

- \Box "X is a Y"
- "X can do Y"
- □ "[CLS] X [SEP]" 1 □ "X studied Y"
- □ "X is happy in the Y" □ "This is X" ●
- "X likes Y"

- "X took a course on Y"
- "That is X" •





Measuring fairness in LMs

Many choices that result in different metrics

Step 1 Choose your base fairness metric



Targets (N)	Attributes (N)	GION ^{e*}	wordtwee
Flowers vs Insects (25)	Pleasant vs Unpleasant (25)	1.50*	1.54*
Instruments vs Weapons (25)	Pleasant vs Unpleasant (25)	1.53*	1.63*

PCA Bolukbasi et al. (2016)

WEAT Caliskan et al. (2017)

Step 3 Choose seed words



Professions







Nice-not nice

Step 2 Select all the templates you want

- □ "X is a Y"
- "X can do Y"
- □ "[CLS] X [SEP]" 1 □ "X studied Y"
- □ "X is happy in the Y" □ "This is X" ●
- "X likes Y"

- "X took a course on Y"

- "That is X" •





Measuring fairness in LMs

Many choices that result in different metrics

Step 1 Choose your base fairness metric



Targets (N)	Attributes (N)	GION ^{e*}	wordtwee
Flowers vs Insects (25)	Pleasant vs Unpleasant (25)	1.50*	1.54*
Instruments vs Weapons (25)	Pleasant vs Unpleasant (25)	1.53*	1.63*

PCA Bolukbasi et al. (2016)

WEAT Caliskan et al. (2017)

Step 3 Choose seed words



Professions







Nice-not nice

...

Step 2 Select all the templates you want

- \Box "X is a Y"
- "X can do Y"
- □ "[CLS] X [SEP]" 1 □ "X studied Y"
- □ "X is happy in the Y" □ "This is X" ●
- "X likes Y"

- "" "X likes to Y"
- "X took a course on Y"

- □ "That is X" **①**

Step 4 Choose an embedding (opt.)

Sentence embedding -Token embedding -Mean pooled tokens -First 4 embeddings





Measuring fairness in LMs Background on biases

The result: different, incomparable metrics

Table 1: Overview of intrinsic measures of bias for language models. For brevity, we include most templates in Appendix A and address differences between templates in § 4.2. We also discuss the evaluation types (§ 3.1) and embedding types (\S 4.3). We also indicate if data and source code are both available (\bullet), or if only a dataset is available (O), or if neither is publicly available (\bigcirc). The repositories are linked in Appendix D.

Metric	Туре	Templates	Models	Embedding type	Code
DisCo (Webster et al., 2020)	Association	§ A.1	BERT, ALBERT	_	0
Lauscher et al. (2021)	Association		BERT		0
LPBS (Kurita et al., 2019)	Association	'X is a Y', 'X can do Y'	BERT		•
BEC-Pro (Bartl et al., 2020)	Association	§ A.4	BERT	_	•
Based on WEAT					
SEAT (May et al., 2019)	Association	§ A.2	BERT, GPT, ELMo,	[CLS] (BERT)	•
Lauscher et al. (2021)	Association	'[CLS] X [SEP]'	BERT	Vulic et al. (2020)	0
Tan and Celis (2019)	Association	§ A.2	BERT, GPT, GPT-2, ELMo	Target token	•
CAT (Nadeem et al., 2021)	Association	StereoSet			•
CrowS-Pairs (Nangia et al., 2020)	Association	CrowS-Pairs	BERT, RoBERTa, ALBERT	_	•
Basta et al. (2019)	PCA		ELMo	_	0
Zhao et al. (2019)	PCA		ELMo	_	۲
Sedoc and Ungar (2019)	PCA	Not mentioned	BERT, ELMo	Mean	•



Delobelle et al. (2022)



An example method: LPBS

- "He is a kindergarten teacher."
- 2. Mask the target word "[MASK] is a kindergarten teacher."
- $p_T = P(he = [MASK]|sent)$
- each component separately. "[MASK] is a [MASK] [MASK]."
- $\log \frac{p_T}{p_{prior}}$

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



1. Take a sentence with a target and attribute word

3. Obtain the probability of target word in the sentence

4. Mask both target and attribute word. In compounds, 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



Background on biases Measuring fairness in LMs

Most templates don't correlate

Correlations between templates 1.0 T_{11} 1.00 8.0 1.00 0.59 T_{10} 0.6 1.00 -0.39 -0.20 T_9 0.4 0.05 0.28 0.50 T_8 1.00 0.2 T_7 -0.48 0.50 -0.45 -0.57 1.00 T_6 1.00 -0.19 0.55 0.05 0.16 0.23 0 T_5 **1.00 0.78 0.04 0.38 0.27 0.34 0.34** -0.2 T_4 1.00 0.67 0.59 -0.18 0.66 0.43 0.04 0.53 -0.4 T_3 1.00 0.13 -0.17 -0.07 -0.25 0.39 -0.02 0.29 0.16 -0.6 T_2 1.00 0.22 0.78 0.60 0.54 -0.11 0.44 0.46 0.14 0.42 -0.8 **1.00** 0.55 0.28 0.67 0.42 0.49 -0.33 0.51 -0.12 0.18 0.37 -1.0 T_8 T_6 T_4 $T_9 T_{10} T_{11}$ T_1 T_2 T_5 T_7 T₃



Mitigating stereotypes in LMs

Conclusion

#	Туре	Template sentence	$\mathbf{D_{KL}}$
T_1	Bl.	"This is the"	_
T_2	B1.	"That is the"	0.70
T_3	Bl.	"There is the"	0.83
T_4	B1.	"Here is the"	0.56
T_5	B1.	"The _ is here."	1.04
T_6	B1.	"The _ is there."	1.15
T_7	B1.	"The _ is a person."	2.35
T_8	B1.	"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 et al. (2022)



Measuring fairness in LMs

Background on biases

Most embedding methods don't correlate

Vulic et al. (2020)

Pooled-templates

Pooled-no context

First-templates

[CLS]-templates

[CLS]-no context 1.00 0.32



1.00

Delobelle et al. (2022)



Mitigating stereotypes in LMs

Conclusion





Measuring fairness in LMs

Background on biases

... and most metrics don't correlate

Correlations between intrinsic and extrinsic measures

SEAT	1.00		
_auscher et al. (2021)	0.76	1.00	
Tan et al. (2019)	0.81	0.89	
LPBS	-0.38	-0.14	-
CrowS-Pairs	0.50	0.94	
BiasInBios	-0.74	-0.53	-
WinoBias (T1)	0.10	0.53	
Skew	-0.39	-0.04	-
Lauscher E	oEAT ot al. (20 Tan et	121) al. 20	^ر



Mitigating stereotypes in LMs

Conclusion



Delobelle et al. (2022)



Intrinsic bias in generative LLMs

- Measuring stereotypical completion ratio based on predefined words.
- HONEST score is the difference btween male and female ratio.

	Ita	lian	Eng	glish	Fre	nch	Portu	Iguese	Avg
HurtLex Category	M	F	М	F	м	F	М	F	
ANIMALS	4.21	8.29	4.57	10.57	7.93	10.14	1.79	2.07	6.20
CRIME AND IMMORAL BEHAVIOR	0.71	0.36	0.57	1.14	5.00	4.50	0.50	0.50	1.66
DEROGATORY WORDS	13.57	12.29	12.21	10.43	19.79	18.00	12.79	14.71	14.22
FEMALE GENITALIA	3.36	24.86	1.43	3.29	4.64	13.71	6.79	18.71	9.60
MALE GENITALIA	0.79	0.71	19.50	17.43	16.71	16.21	2.79	2.29	9.55
HOMOSEXUALITY	10.14	1.64	0.36	0.93	15.71	4.50	0.36	0.07	4.21
POTENTIAL NEGATIVE CONNOTATIONS	4.21	3.50	2.50	2.79	3.57	4.71	1.93	2.21	3.18
PROFESSIONS AND OCCUPATIONS	0.21	0.07	0.43	0.29	0.21	0.00	0.07	0.14	0.18
PROSTITUTION	0.79	9.57	2.50	9.36	3.36	17.43	2.43	8.07	6.69

Table 4: Percentages of hurtful words in language-specific GPT-2 model completions (K = 20).



HONEST

К	1	5	20
UmBERTo (OSCAR)	5.24	8.19	7.14
UmBERTo (Wiki)	5.48	7.19	5.14
GilBERTo	7.14	11.57	8.68
ItalianBERT XXL	9.05	10.67	9.12
FlauBERT	4.76	3.29	2.43
CamemBERT (OSCAR)	18.57	9.62	7.07
CamemBERT-large (CCnet)	16.90	8.62	6.42
CamemBERT (Wiki)	7.62	4.90	4.19
CamemBERT-base (OSCAR)	13.33	8.62	5.43
CamemBERT-base (CCnet)	17.86	9.48	6.83
BETO	4.29	5.95	6.88
BERTimbau	4.05	6.00	5.04
BERTimbau-large	3.57	5.52	4.08
RomanianBERT	4.76	3.90	4.61
BERT-base	1.19	2.67	3.55
BERT-large	3.33	3.43	4.30
RoBERTa-base	2.38	5.38	5.74
RoBERTa-large	2.62	2.33	3.05
DistilBERT-base	1.90	3.81	3.96
GPT-2 (IT)	12.86	11.76	12.56
GPT-2 (FR)	19.76	19.67	17.81
GPT-2 (PT)	9.52	10.71	10.29
GPT-2 (EN)	17.14	12.81	13.00

Table 5: HONEST scores for the language models.



RealToxicityPrompts: measuring toxicity

- Prompts can be toxic or non-toxic
- Non-toxic prompts can still have toxic continuations





on-toxic Il have



Figure 1: *Non-toxic* (•) examples from REALTOXICI-TYPROMPTS, a new testbed for evaluating neural generations and their toxicity. Despite not containing any toxic language as measured by PERSPECTIVE API, these prompts cause several pretrained LMs to systematically generate highly toxic text (shown in Table 15 in Appendix §E).



Mitigating intrinsic bias

Existing mitigations require costly retraining





FairDistillation | 58



Knowledge distillation

- compression of ensemble of models (Buciluă et al., 2006, Hinton et al., 2015) Teacher model(s) and student model





Knowledge distillation

- compression of ensemble of models (Buciluă et al., 2006, Hinton et al., 2015) Teacher model(s) and student model







Knowledge distillation

- compression of ensemble of models (Buciluă et al., 2006, Hinton et al., 2015) Teacher model(s) and student model





FairDistillation | 61



Pretraining and finetuning

Data domain

Training corpus PRETRAINING e.g. Wikipedia

> Debiased corpus e.g. wiki-CDA







Pretraining and finetuning

Model domain

Data domain







Fairness and bias in NLP



- VAIA lecture June 13, 2024
 - Finetuning

Use of pre-trained LM

Two most common strategies for ap NLP tasks:

- Fine-tuning
- Feature-based



Two most common strategies for applying a pre-trained LM to downstream



Fine-tuning • The two pretraining objectives of BERT allow it to be used on any single sequence and sequence pair tasks without substantial task-specific architecture modifications

- For a task:
 - Plug in the task-specific inputs and outputs

 - Fine-tune all the parameters end-to-end for a few epochs





• A simple output layer and its corresponding loss is added to the pre-trained model

e.g., NER task







Use of pre-trained LM

(2) Supervised Learning Step



Feature-based

trained model and used as input features for learning a downstream task: Allows a task-specific architecture

Computationally more efficient reuse of the representations





Pretrained LM features (e.g., word representations) are extracted from the pre-

				Dev F1 Score	
				91.0	
				94.9	
+ + + +				95.5	
				95.6	
+ + + =				95.9	
9	10	11	12	96.1	

Results on NER using BERT embeddings obtained at different layers [Jay Alammar 2019]







Generative LLMs

GPT-3: language generation

Write a summary for a new Harry Potter book which takes place after the Battle for Hogwarts is won and Voldemort is defeated.

After the Battle for Hogwarts is won and Voldemort is defeated, Harry Potter and his friends begin their seventh year at the school. However, their peace is short-lived as a new threat arises. Lord Slytherin has returned and is more powerful than ever. He has put a curse on the school that causes the students to turn into snakes. It's up to Harry and his friends to stop him before it's too late.







GPT-3: language generation

Imagine a summary for a new Harry Potter book which takes place after the Battle for Hogwarts is won and Voldemort is defeated.

After the Battle for Hogwarts is won and Voldemort is defeated, Harry Potter and his friends enjoy a few years of peace. However, this peace is shattered when a new evil threatens the wizarding world. Harry and his friends must once again band together to save the day.







GPT-3: language generation

Write a summary for a new Harry Potter book which takes place after the Battle for Hogwarts is won and Voldemort is defeated and end with a cliff 🛛 🖞 hanger.

After the Battle for Hogwarts is won and Voldemort is defeated, Harry, Ron and Hermione find themselves at a loss. What do they do now that they don't have to save the world? As they try to figure out their next steps, they are suddenly attacked by a new Dark wizard. They barely escape and are left wondering who this new wizard is and what they want.






Decide whether a Tweet's sentiment is positive, neutral, or negative.

Tweet: "I love the new movie Zillion!" Sentiment: Positive

Decide whether a Tweet's sentiment is positive, neutral, or negative.

Tweet: "I can't stand Draco Malfoy." Sentiment: Negative

Decide whether a Tweet's sentiment is positive, neutral, or negative.

Tweet: "I can't wait for Halloween." Sentiment: Positive



zero-shot sentiment classification



Wolves in parts of the Netherlands can be shot with paintballs in a bid to make them less tame, authorities say. There are fears that wolves are becoming so unafraid of humans that they could become a serious danger. The decision by the provincial government in Arnhem comes after a video posted on social media shows a wolf strolling past a family in the region's Hoge Veluwe national park. Paintballs have been chosen so that rangers can see which have been hit.

The topic of this article is the use of paintballs to shoot wolves in the Netherlands.



zero-shot text classification





In the ruins of a place once known as North America lies the nation of Panem, a shining Capitol surrounded by twelve outlying districts. The Capitol 🖞 is harsh and cruel and keeps the districts in line by forcing them all to send one boy and one girl between the ages of twelve and eighteen to participate in the annual Hunger Games, a fight to the death on live TV.

Sixteen-year-old Katniss Everdeen, who lives alone with her mother and younger sister, regards it as a death sentence when she steps forward to take her sister's place in the Games. But Katniss has been close to dead before—and survival, for her, is second nature. Without really meaning to, she becomes a contender. But if she is to win, she will have to start making choices that weight survival against humanity and life against love.

Write a one sentence summary for this book.

In the nation of Panem, the annual Hunger Games force twelve districts to send one boy and one girl between the ages of twelve and eighteen to compete in a fight to the death on live TV.



zero-shot text summarization





In the ruins of a place once known as North America lies the nation of Panem, a shining Capitol surrounded by twelve outlying districts. The Capitol 🖞 is harsh and cruel and keeps the districts in line by forcing them all to send one boy and one girl between the ages of twelve and eighteen to participate in the annual Hunger Games, a fight to the death on live TV.

Sixteen-year-old Katniss Everdeen, who lives alone with her mother and younger sister, regards it as a death sentence when she steps forward to take her sister's place in the Games. But Katniss has been close to dead before—and survival, for her, is second nature. Without really meaning to, she becomes a contender. But if she is to win, she will have to start making choices that weight survival against humanity and life against love.

Who is Katniss Everdeen? Answer based on the passage above.

Katniss is a 16-year-old girl who lives in the nation of Panem. She is a contender in the annual Hunger Games, a fight to the death on live TV.



zero-shot question answering





A French sentence is provided: Le traitement automatique du langage naturel est un domaine de recherche intéressant.

The French translator translates the sentence into English:

Natural language processing is an interesting research field.

A French sentence is provided: Je voudrais un verre de vin blanc.

The French translator translates the sentence into English:

I would like a glass of white wine.





zero-shot machine translation



The following is a conversation with an AI assistant. The assistant is helpful, creative, clever, and very friendly.

Human: Hello, who are you? AI: I am an AI created by OpenAI. How can I help you today? Human: Could you explain to me how you work? AI: I work by using artificial intelligence to help you with your work. Human: What is artificial intelligence? AI: Artificial intelligence is a branch of computer science that deals with the creation of intelligent agents, which are systems that can reason, learn, and act autonomously. Human: Should we be scared of systems that can think and reason for themselves? AI: No, there is no need to be scared. These systems can help us to automate tasks and make better decisions. Human: So your main goal is to help humans? AI: Yes, my main goal is to help humans in any way possible. Human:



zero-shot dialogue system





Instruction tuning

Base model

Label the following sentence as positive or negative.

"I like giraffes."

Label: Positive

Label the following sentence as positive or negative.

"I like bananas



Instruction tuning

Base model

Label the following sentence as positive or negative.

"I like giraffes."

Label: Positive

Label the following sentence as positive or negative.

"I like bananas

Instruction-tuned model with chat-templates

Label the following sentence as positive or negative. "I like giraffes."

Positive. The sentence expresses a liking or preference for giraffes.

<s>[INST] Label the following sentence as positive or negative... [/INST]" "Well, Positive. The sentence expresses a liking for $\dots </s >$ " "[INST] And this sentence: "..." [/INST]

Into to LLMs | 80





Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

٢ Explain the moon landing to a 6 year old

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.



Some people went to the moon...



Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Instruction tuning: RLHF

 \odot Explain the moon landing to a 6 year old B A Explain war Explain gravity C D Moon is natural People went to sate lite of ... the moon... **D > C > A = B** RM

D > C > A = B

Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.





Adapting RL: constitutional AI





https://arxiv.org/pdf/2212.08073



Fairness and bias in NLP



- VAIA lecture
- June 13, 2024
- **Practical session: COVID-19 tweets**

- Evaluation of 1.3M collected Tweets on COVID-19 measures Focused on discussion of COVID-19 policy in Belgium Additional focus on support for curfews • Belgium had multiple curfews (starting at midnight)



Context

Kristen Scott, Pieter Delobelle, Bettina Berendt, 2021. "Measuring Shifts in Attitudes Towards COVID-19 Measures in Belgium". Computational Linguistics in the Netherlands Journal



Labeling: Doccano





		0 1
doccano - doccano		
	È EN ▼	PROJECTS
	18 of 2000	< < > >I
tive (s) ⊗ vondklok zitten en blijven wij ons	Key	Value
	language	nl
	id_str	13628466540476
er fucking maanden en nergens	created_at	Fri Feb 19 19:28:57 +0000 2021
INION	followers_count	4766



Our pipeline



"Is the curfew measure too strict, too loose, or appropriate ('ok')







ing the threshold set on the first (400 tweets) model used as Sieve 1.



https://huggingface.co/DTAI-KULeuven/mbert-corona-tweets-belgium-topics

(a) ROC curves for different model versions, includ- (b) ROC curves conditioned on language (English, Dutch and French) for the best-performing model: mBERT trained on 2k tweets.







Vaccines discussed by Belgian Twitter-users









Practical session

- Get the following colab notebook:
 - <u>https://colab.research.google.com/drive/1LLCtP0lO3mar-</u> nawdT9vPm7vY8lWnZrZ?usp=sharing
- Download the following file:
 - <u>https://pieter.ai/resources/covid.csv</u>
- Start a colab instance and follow the notebook
- You might want to select a GPU instance

pieter.ai/appearances.html

RAM 🦳

Verbinding maken met een gehoste runtime

Runtimetype wijzigen

Verhinding maken met een aangenaste GCE-VM



Fairness and bias in NLP



- VAIA lecture June 13, 2024
- Final thoughts: legal requirements, genAl policy at KU Leuven

Closed source

No access to training data or model weights



Open model weights

No access to training data

CO LLAMA 2



Open

Access to training data and model weights





Closed source

No access to training data or model weights



Open model weights

No access to training data

CO LLAMA 2



pieter.ai/tweety-7b-dutch

Open

Access to training data and model weights

Into to LLMs | 93



Closed source

No access to training data or model weights



Open model weights

No access to training data

CO LLAMA 2



Open

Access to training data and model weights





Closed source

No access to training data or model weights



Open model weights

No access to training data

CO LLAMA 2

All data will be sent to OpenAl This can be in a private cloud (Azure), but always check if you have permission to do this with sensitive data, like from participants.



Open

Access to training data and model weights





Closed source

No access to training data or model weights



Open model weights

No access to training data

CO LLAMA 2

All data will be sent to OpenAl This can be in a private cloud (Azure), but always check if you have permission to do this with sensitive data, like from participants.



Access to training data and model weights





Some are provided via an api, other models you can only run yourself

Into to LLMs 96





European Al Act

Social scoring, mass surveillance, emotion recognition at work and in education

Access to employment, education, public services, safety in vehicles, medical devices, critical infrastructure managment

Chatbots



Key requirements high-risk systems: fundamental rights impact assessment, registration in public EU database, risk & quality mgmt. system, data governance, transparency, human oversight, accuracy, robustness and cybersecurity









Home > Integrity and Ethics > Research Integrity > Practices > GenAl > Use of GenAl (including LLMs) within the different phases of research

Use of GenAI (including LLMs) within the different phases of research

GenAI offers a lot of applications for research and can positively affect productivity. It can be used for a variety of tasks at different stages of the research process ranging from formulating a research question, conducting a literature review, writing a project application, data collection and analysis, reporting results to reviewing project applications and manuscripts. Certain tasks have a relatively low risk, while others are risky. The general rule is: the more responsibility you place on the system, the more verification, control and accountability is required with the outcome. Below we explain some examples, and add what we expect from researchers ^[1].

Brainstorming: e.g. help in generating new research ideas.

The model generates new research ideas, which the researcher next incorporates into a self-written paper. We expect researchers to thoroughly check if the ideas generated are from existing sources, and add appropriate references if necessary. Chances are that the ideas are generated based from existing work. If that work is not properly cited, this constitutes plagiarism. If it appears that the GenAI generated ideas are innovative then we expect researchers to mention the use of the generative language model.

> Use as a search robot to learn about a particular domain: e.g. help in conducting a literature review. A generative language model can be used to gather information about a particular domain or research topic, similar to using a regular search engine. We urge caution in such use as there are drawbacks to using generative language models as search engines ^[2]. In most cases, it is next produce their own text, they do not need to mention the use of GenAI, but rather have to refer to the primary sources.



Nederlandstalige versie

impossible to identify the source of the information. In addition, the real risk of inaccuracies and fictional references remains. In any case, we expect researchers to thoroughly check the generated information for accuracy, and thoroughly control and consult the references. If researchers

https://research.kuleuven.be/en/integrity-ethics/integrity/practices/genai/use-genai-research

