LLM Course

LLM Course documentation

Bias and limitations

LLM Course

■ View all resources

Agents Course

Audio Course

Community Computer Vision Course

Deep RL Course

Diffusion Course

LLM Course

MCP Course

ML for 3D Course

ML for Games Course

Open-Source AI Cookbook

Robotics Course

a smol course

Search documentation

AR

BN

DE

EN

ES

FA

FR

GJ

HE

HI

ID

IT

JA

KO

MY

NE

PL

PT

RO

RU

RUM

TE

TH

TR

VI

ZH-CN

ZH-TW

Join the Hugging Face community

and get access to the augmented documentation experience

Collaborate on models, datasets and Spaces

Faster examples with accelerated inference

Switch between documentation themes

Sign Up

to get started

Copy page

Bias and limitations

If your intent is to use a pretrained model or a fine-tuned version in production, please be aware that, while these models are powerful tools, they come with limitations. The biggest of these is that, to enable pretraining on large amounts of data, researchers often scrape all the content they can find, taking the best as well as the worst of what is available on the internet.

To give a quick illustration, let’s go back to the example of a

fill-mask

pipeline with the BERT model:

Copied

from

transformers

import

pipeline

unmasker = pipeline(

"fill-mask"

, model=

"bert-base-uncased"

)

result = unmasker(

"This man works as a [MASK]."

)

print

([r[

"token_str"

]

for

r

in

result])

result = unmasker(

"This woman works as a [MASK]."

)

print

([r[

"token_str"

]

for

r

in

result])

Copied

[

'lawyer'

,

'carpenter'

,

'doctor'

,

'waiter'

,

'mechanic'

]

[

'nurse'

,

'waitress'

,

'teacher'

,

'maid'

,

'prostitute'

]

When asked to fill in the missing word in these two sentences, the model gives only one gender-free answer (waiter/waitress). The others are work occupations usually associated with one specific gender — and yes, prostitute ended up in the top 5 possibilities the model associates with “woman” and “work.” This happens even though BERT is one of the rare Transformer models not built by scraping data from all over the internet, but rather using apparently neutral data (it’s trained on the

English Wikipedia

and

BookCorpus

datasets).

When you use these tools, you therefore need to keep in the back of your mind that the original model you are using could very easily generate sexist, racist, or homophobic content. Fine-tuning the model on your data won’t make this intrinsic bias disappear.

Update

on GitHub

←

Inference with LLMs

Summary

→