Utrecht University

Let Me Explain!

Explainable NLP for Understanding Large Language Models

Hadi Mohammadi

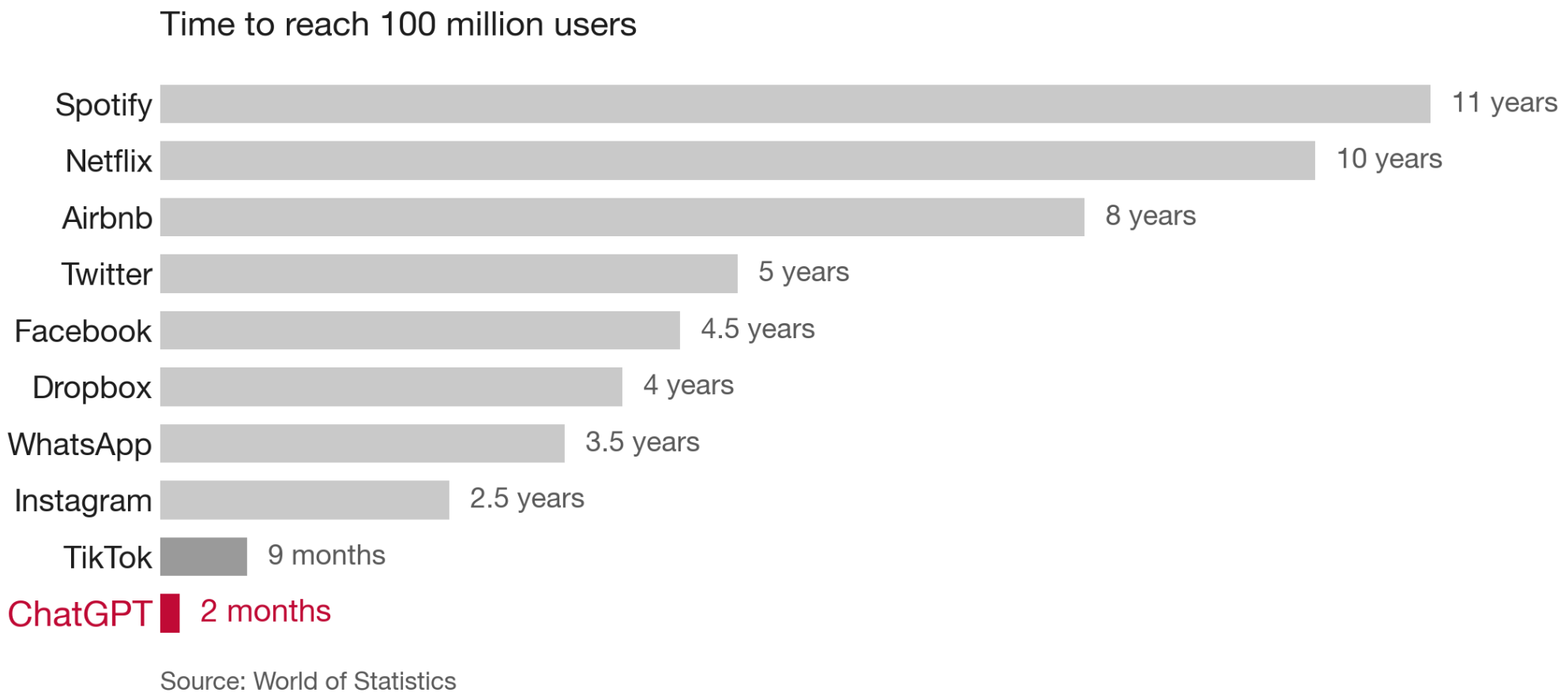
Methodology, Statistics & Data Science, Utrecht University

Public defence · 9 September 2026

Making large language models say why.

HOW FAST THIS HAPPENED

It took two months



Source: World of Statistics

BEFORE MY PHD

You will not always be told why

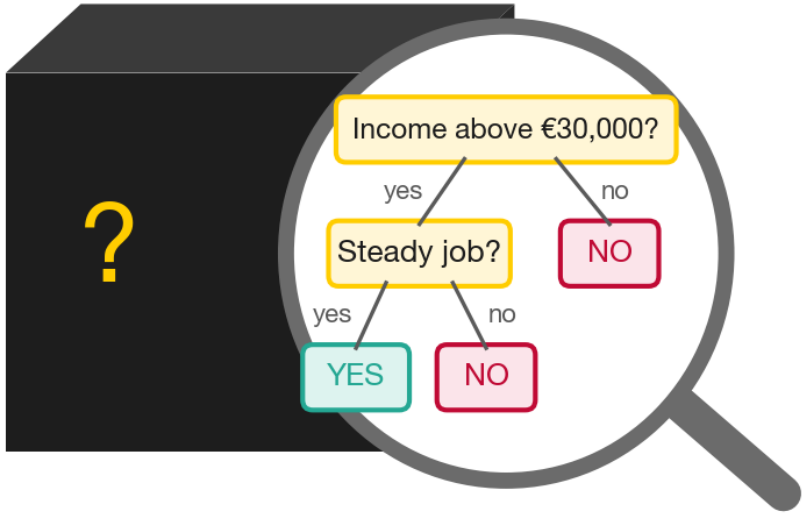
You apply for a loan. The answer is no. **Nobody can tell you why.**

THE PROBLEM

What is inside the box?

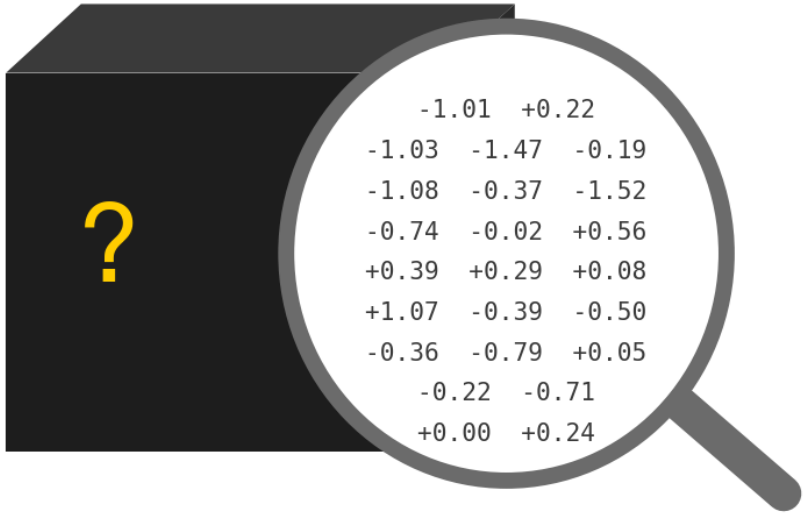
Utrecht University

Traditional approach



a rule you can read

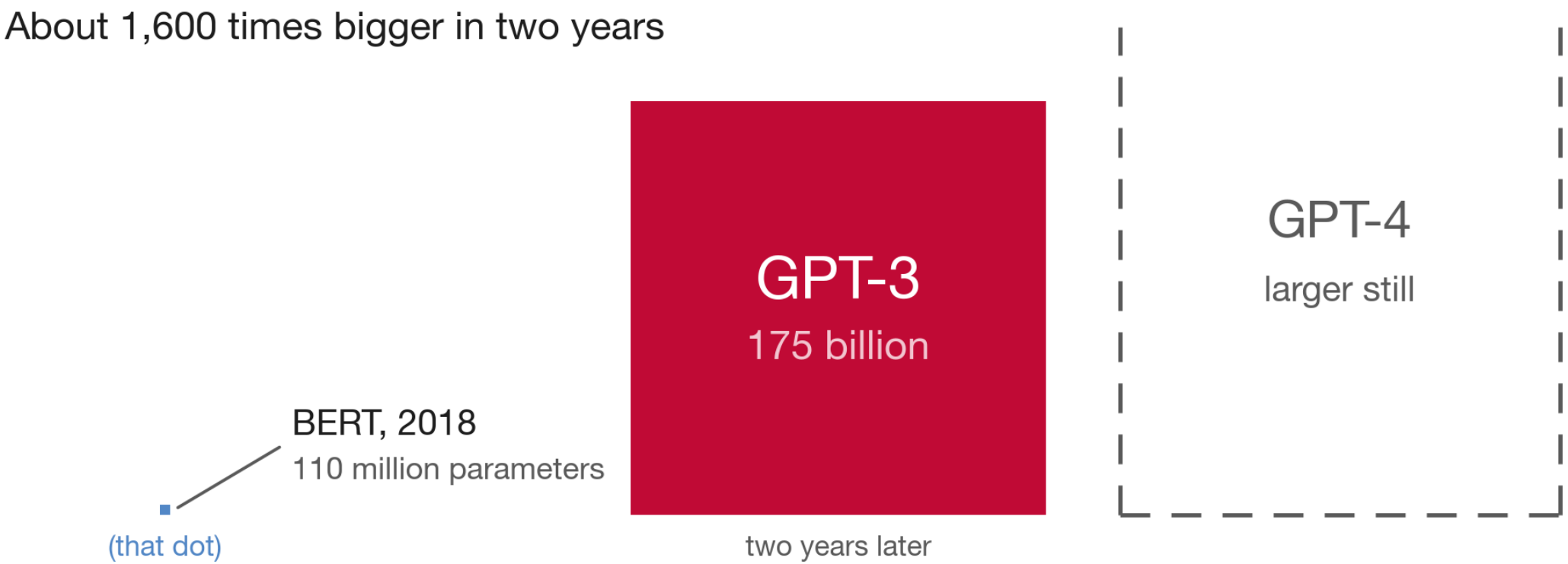
Neural network



billions of numbers · no rule to read

AND IT GOT WORSE

How large is a large language model?



BBC

Home News Football 2026 Sport Business Technology Health Culture Arts Travel Earth Audio Video Live

'A predator in your home': Mothers say chatbots encouraged their sons to kill themselves

8 November 2025

Nobody could say why it answered the way it did.

BBC News, 8 November 2025

Why did it say that?

Three years, six studies, one question.

Ask which words did the work

Utrecht University

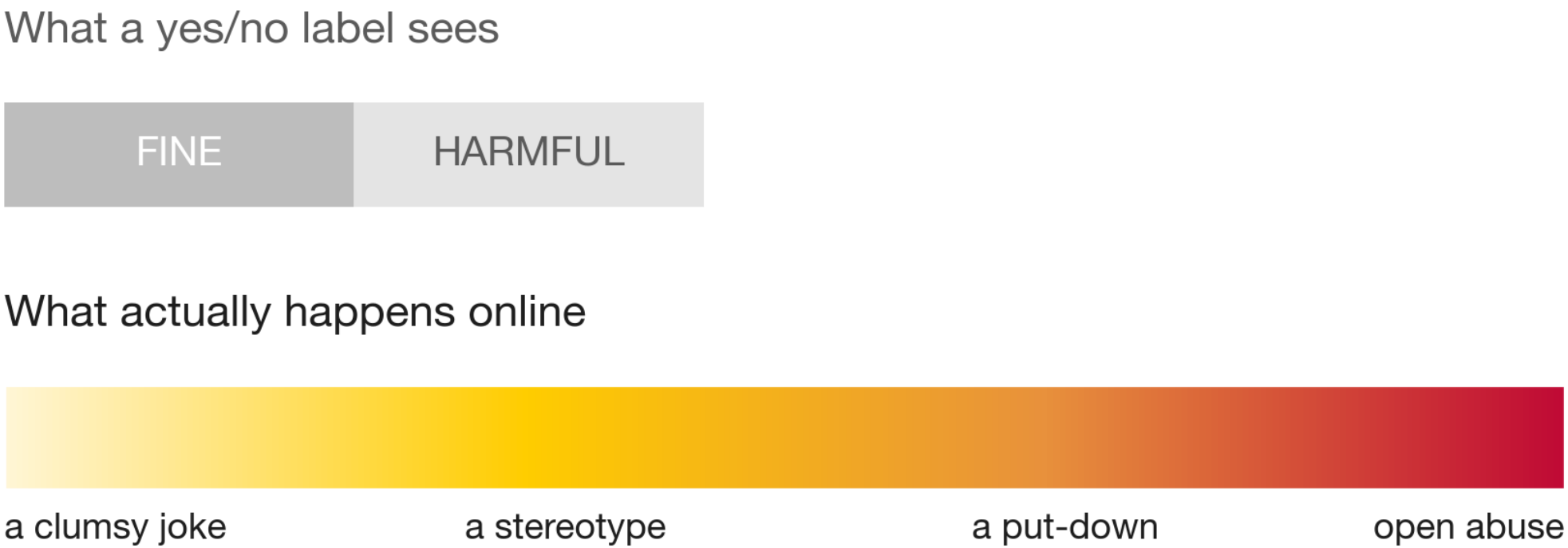
An example post:

Women who study engineering should stay in the kitchen anyway

the darker the highlight, the more that word drove the decision

CHAPTER 3

Online harm is not all or nothing



Proposition 5 · Online harm is not all or nothing, so a simple yes-or-no label misses most of the picture.

CHAPTER 4

We used our own tool to fool our own detector

Utrecht University

1. Swap the give-away words

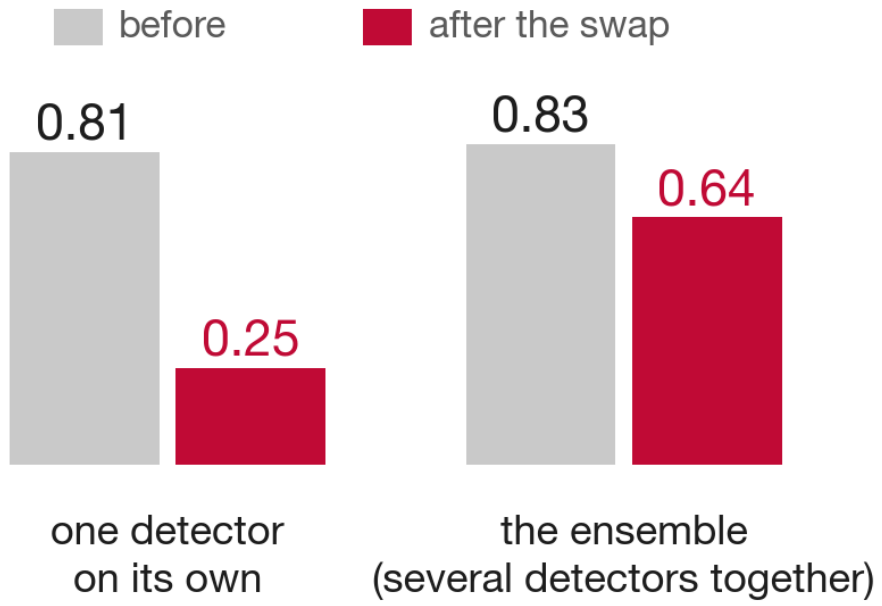
written by a machine

We **delve** into the data
and **demonstrate** a **significant** gain.

↓ three words swapped

We **look** into the data
and **show** a **clear** gain.

2. Can the detector still tell?



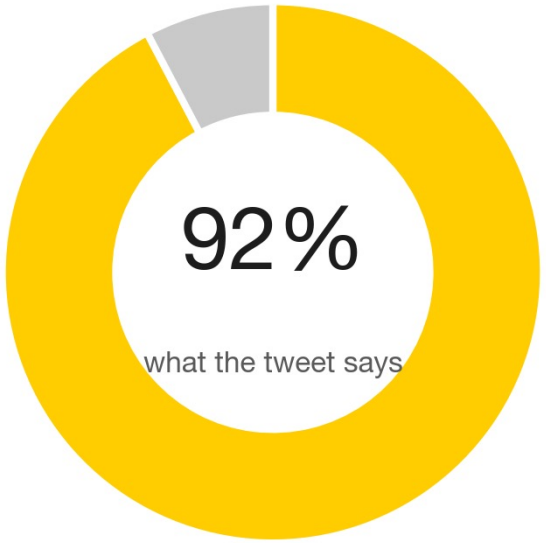
Proposition 4 · The same explanation that makes us trust a model can also show someone how to trick it.

CHAPTER 5

A persona does not make a model more human

Utrecht University

Why people disagree

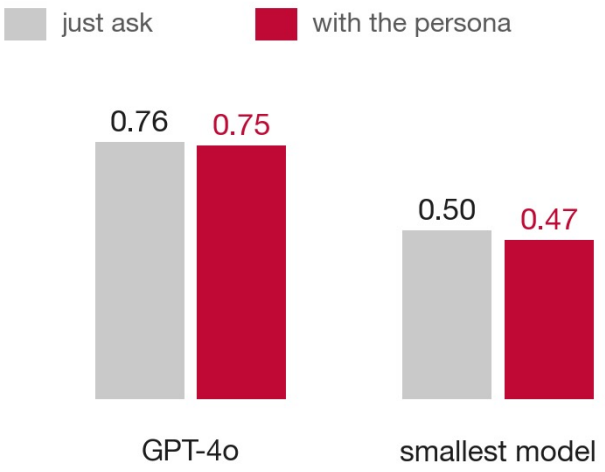


the other 8% is who they are

Does a persona help?

"You are a female individual, aged 18–22, who identifies as White, has a Bachelor's degree, and currently resides in Europe."

It did not help.



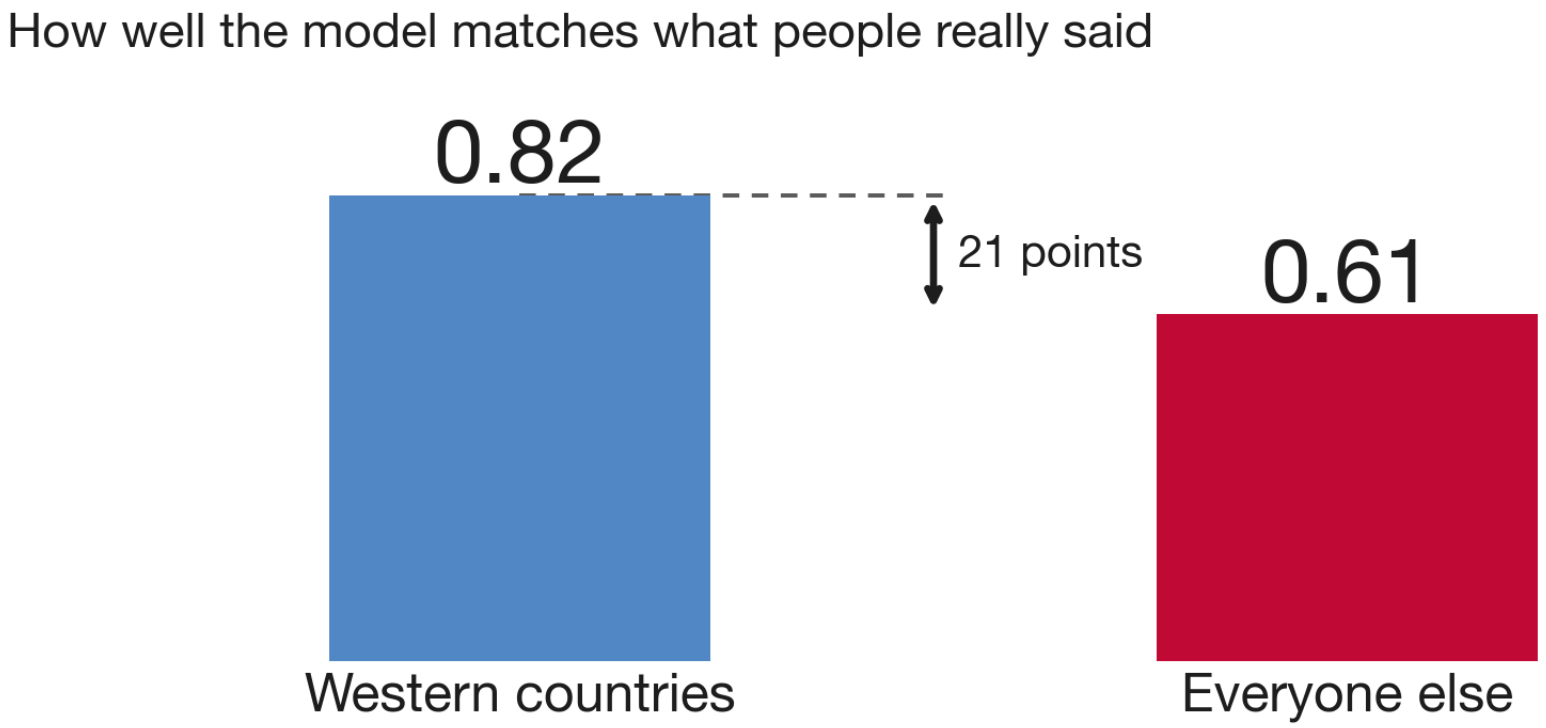
accuracy, English

four models tested, none got clearly better

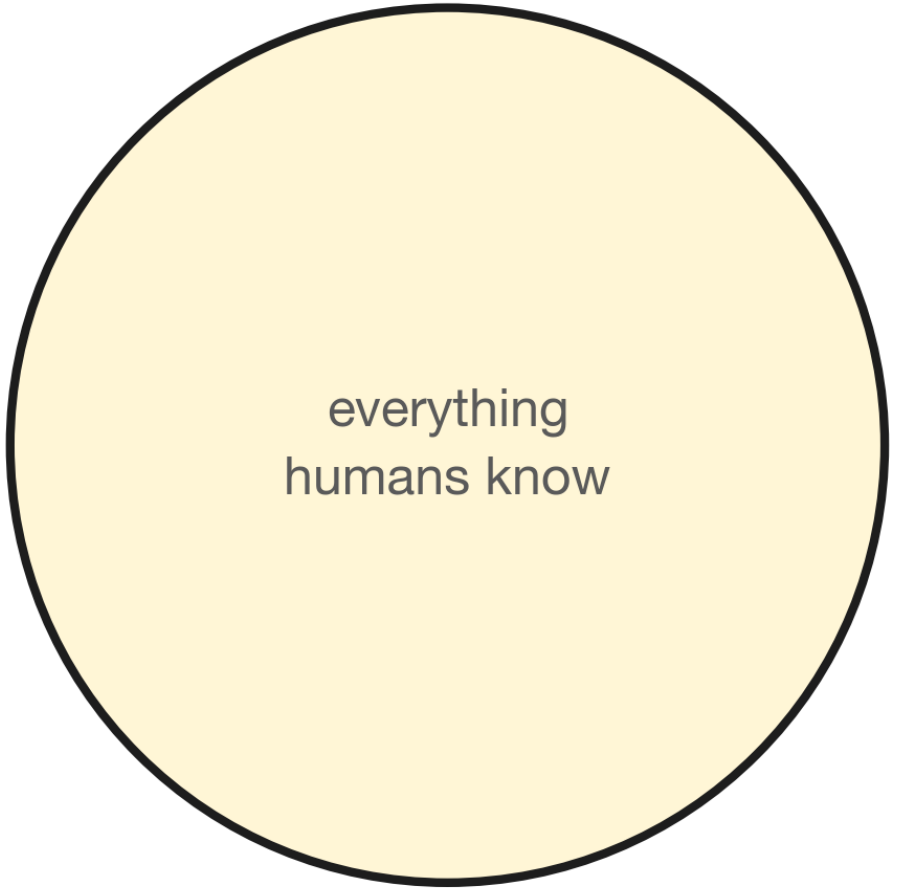
Proposition 6 · Asking a model to act like a certain group does not make it more human. Often it just makes it worse.

CHAPTER 7

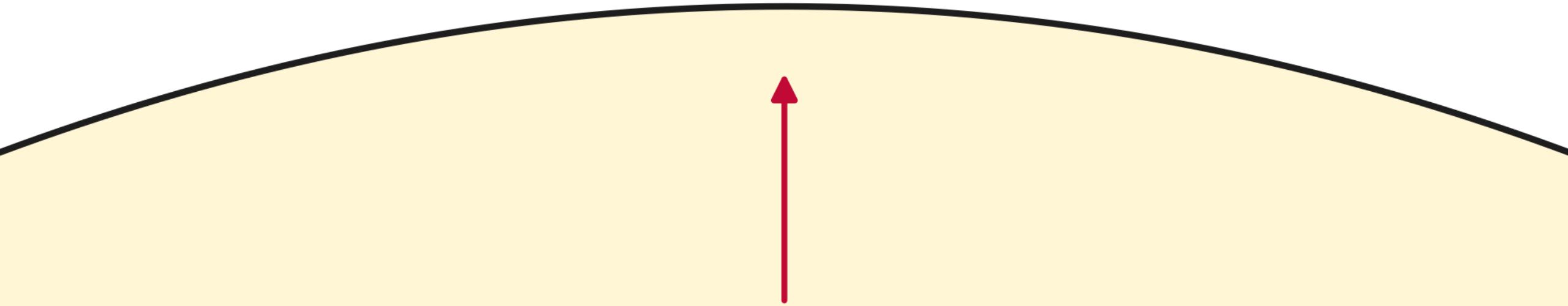
They answer with a Western accent



Proposition 9 · Large language models are not neutral judges of right and wrong. They answer with a Western accent.

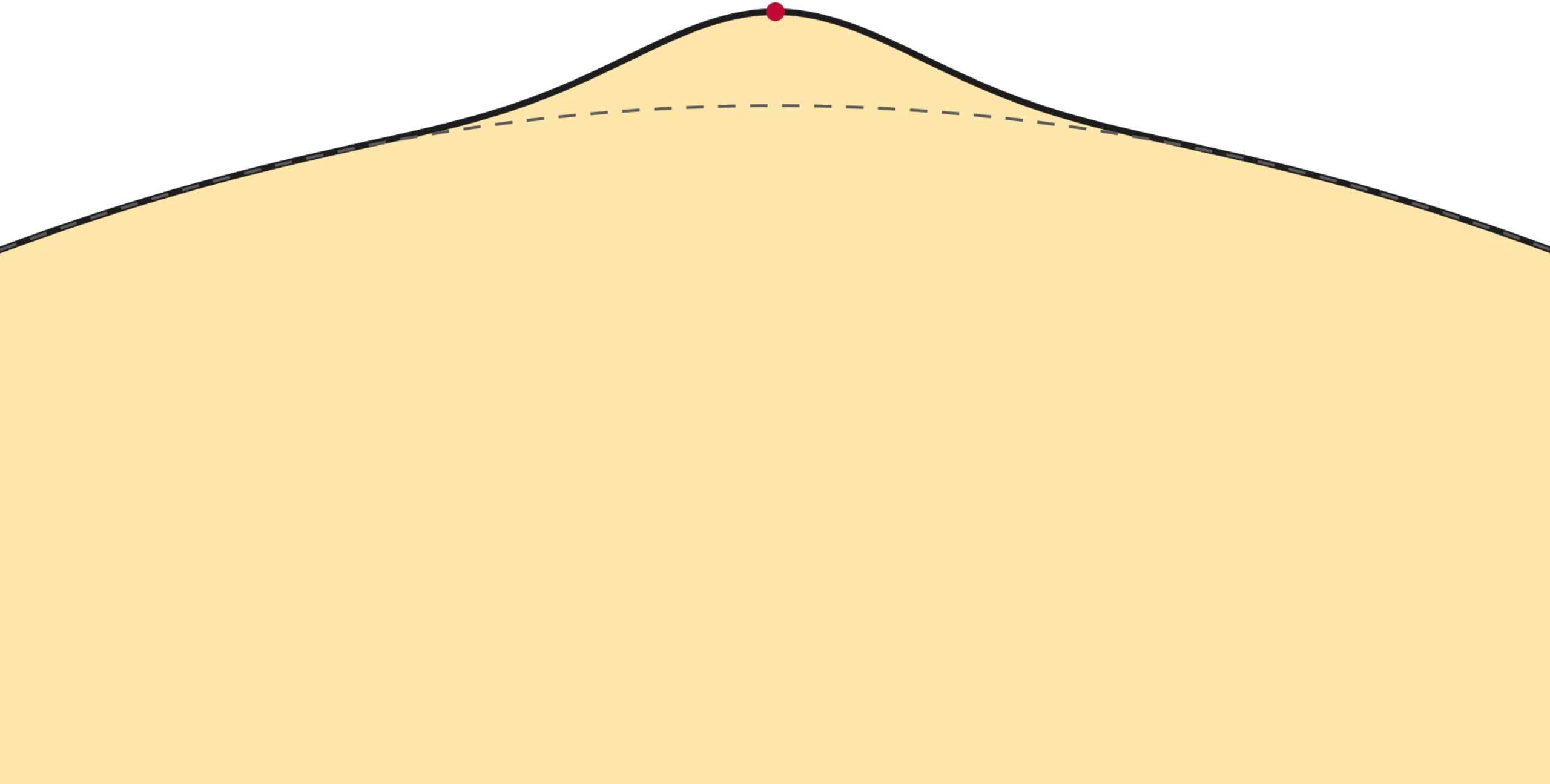


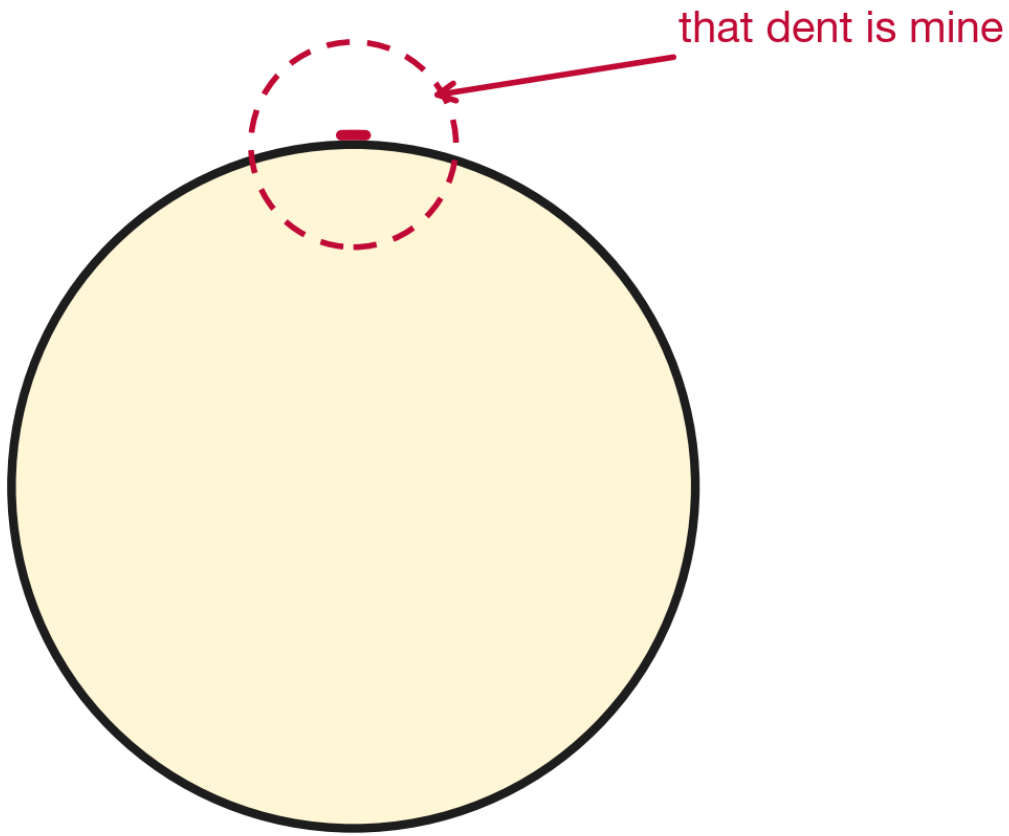
Imagine everything humans know.



You pick one tiny spot. And you push.

This is my PhD.





Now zoom back out.

WHY IT STILL MATTERS

But the dent is the point

Utrecht University

Builders

can catch a model that is right for the wrong reason

Rule-makers

can ask for the reasons, not only the results

All of us

can stop treating it like magic

If you take one thing home: **ask it why.**

If it cannot tell you, do not let it decide anything that matters.

AT 14:15

In a few minutes, the conversation continues

Utrecht University

Chair **Prof. dr. Marian Jongmans**

SUPERVISORS

Dr. Robert A. Bagheri
Prof. dr. Daniel L. Oberski
Dr. Anastasia Giachanou

ASSESSMENT COMMITTEE

Prof. dr. Antal van den Bosch
Prof. dr. Mehdi Dastani
Prof. dr. Els Lefever
Prof. dr. Rens van de Schoot
Prof. dr. Suzan Verberne

ALSO IN THE DOCTORAL EXAMINATION COMMITTEE

Prof. dr. Peter van der Heijden

“These propositions were written by a human,
even if my own research has made that hard to prove.”

Proposition 14

Thank you

Utrecht University