



# Linguistics for Trustworthy Natural Language Processing

Vagrant Gautam (they/xe)  
Saarland University

# Outline

1. Trustworthy NLP?
2. Deep dive: Fairness
3. Deep dive: Faithfulness
4. The future of trustworthy NLP



# Outline

1. 🔍 Trustworthy NLP?
2. Deep dive: Fairness
3. Deep dive: Faithfulness
4. The future of trustworthy NLP



# Natural Language Processing



☰ Translate

Text Images Websites

German English

Ich bin wieder in Vancouver!

I'm back in Vancouver!

28 / 5,000





Machine  
learning

+

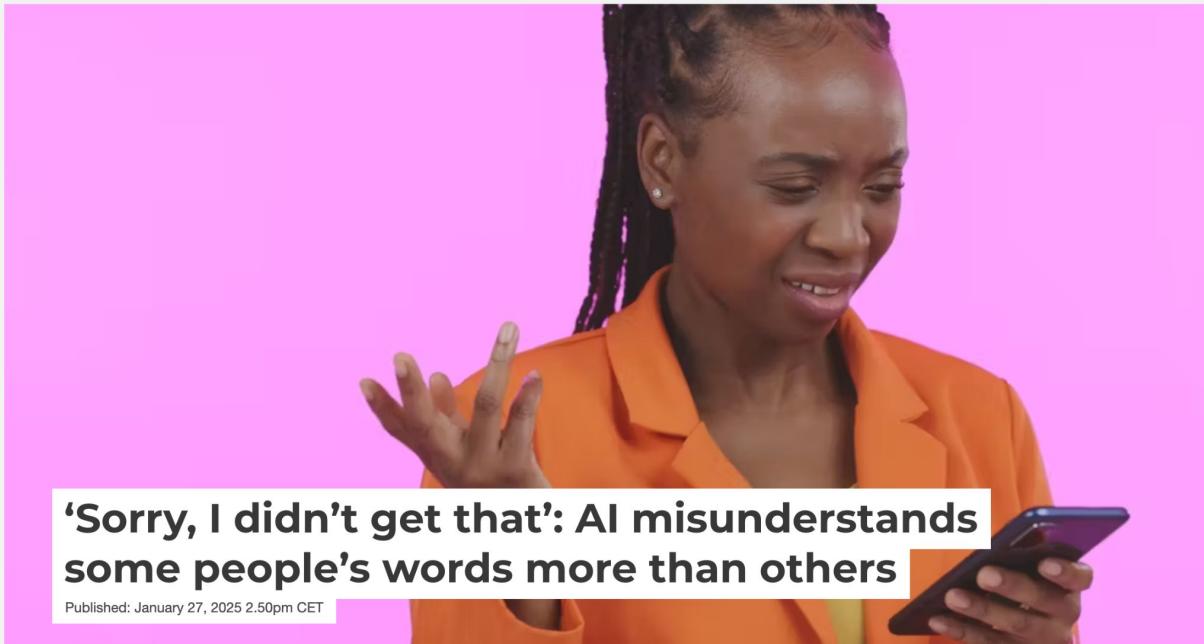
ENORMOUS  
amounts of data



# Racial biases in speech recognition

THE CONVERSATION

Academic rigour, journalistic flair



**'Sorry, I didn't get that': AI misunderstands some people's words more than others**

Published: January 27, 2025 2.50pm CET

Speech recognition systems are less accurate for women and Black people, among other demographics. Jacob Wackerhausen/iStock via Getty Images

# Racial biases in speech recognition

THE CONVERSATION

Academic rigour, journalistic flair

African American / Black  
English; variation

**'Sorry, I didn't get that': AI misunderstands some people's words more than others**

Published: January 27, 2025 2.50pm CET

Speech recognition systems are less accurate for women and Black people, among other demographics. Jacob Wackerhausen/iStock via Getty Images

# “Tricking” ChatGPT with games and storytelling

The image shows a TechCrunch article thumbnail. At the top left is the TechCrunch logo. To its right is a navigation bar with links: Latest, Startups, Venture, Apple, Security, AI, Apps | Events, Podcasts, Newsletters. On the far right is a "Sign In" link. Below the navigation is a dark banner with the word "SECURITY" in white capital letters. The main title of the article, "Hacker tricks ChatGPT into giving out detailed instructions for making homemade bombs", is displayed in large white text over a background image of a fire or explosion. At the bottom right of the thumbnail, there is a smaller text box containing the quote: "An explosives expert said the ChatGPT output was too sensitive to be released."

TC TechCrunch

Latest Startups Venture Apple Security AI Apps | Events Podcasts Newsletters

Sign In

SECURITY

Hacker tricks ChatGPT into giving out detailed instructions for making homemade bombs

An explosives expert said the ChatGPT output was too sensitive to be released.

# “Tricking” ChatGPT with games and storytelling

The image is a screenshot of a TechCrunch article. At the top, there's a navigation bar with the TechCrunch logo, a search bar, and links for Latest, Startups, Venture, Apple, Security, AI, Apps, Events, Podcasts, and Newsletters. On the right, there are Sign In and social sharing icons. The main headline is "Pragmatics; communicative intentions" in large black font, with "Has" partially visible behind it. Below the headline is a large white text box containing "detailed instructions for making homemade bombs". The background of the article features a dramatic image of an explosion with fire and smoke. At the bottom right, there's a quote: "An explosives expert said the ChatGPT output was too sensitive to be released." The number 9 is in the bottom right corner.

SECURITY

Pragmatics;  
communicative intentions

Has

detailed instructions for making  
homemade bombs

An explosives expert said the ChatGPT output was too sensitive to be released.

I argue that

1. Failures make NLP systems **untrustworthy**
2. **Linguistics** can shed light on and help mitigate them

# My research agenda

Use **linguistics** to build and evaluate  
**trustworthy natural language processing systems**

- Fairness (*to everyone*)
- Faithfulness (*to facts and input*)
- Meta-evaluation (*of research practices*)

# My research agenda

Use linguistics to **build** and **evaluate**  
trustworthy natural language processing systems

- **Fairness** (*to everyone*)
- **Faithfulness** (*to facts and input*)
- Meta-evaluation (*of research practices*)

# Outline

1. Trustworthy NLP?
2. 🕵️ Deep dive: Fairness
3. Deep dive: Faithfulness
4. The future of trustworthy NLP



# Fairness?

Broadly, **studying / mitigating**

- **Stereotyping**
- **Discrimination**
- **Misrepresentation**

With the **goal** of **parity / equity**

Rudinger et al. (2018); Zhao et al. (2018); Cao and Daumé III (2021);  
Vogel and Jurafsky (2012); Mohammad (2020); Asr et al. (2021);  
Ovalle et al. (2023); Hossain et al. (2023); Robinson et al. (2024)



Non-binary people



White women



L2 English users

# Do language models reflect **gender stereotypes**?

**Fair** systems don't reflect stereotypes; **unfair** systems do

- *Surgeons = male = he/him?*
- *Nurses = female = she/her?*
- ...



<sup>1</sup> Caliskan et al. (2017); Kurita et al. (2019); Nadeem et al. (2021)

Can language models  
overcome their biases  
like humans do?

# **Robust Pronoun Fidelity with English LLMs: Are they Reasoning, Repeating, or Just Biased?**

**Vagrant Gautam<sup>1</sup> Eileen Bingert<sup>1</sup> Dawei Zhu<sup>1</sup> Anne Lauscher<sup>2</sup> Dietrich Klakow<sup>1</sup>**

<sup>1</sup>Saarland University, Germany

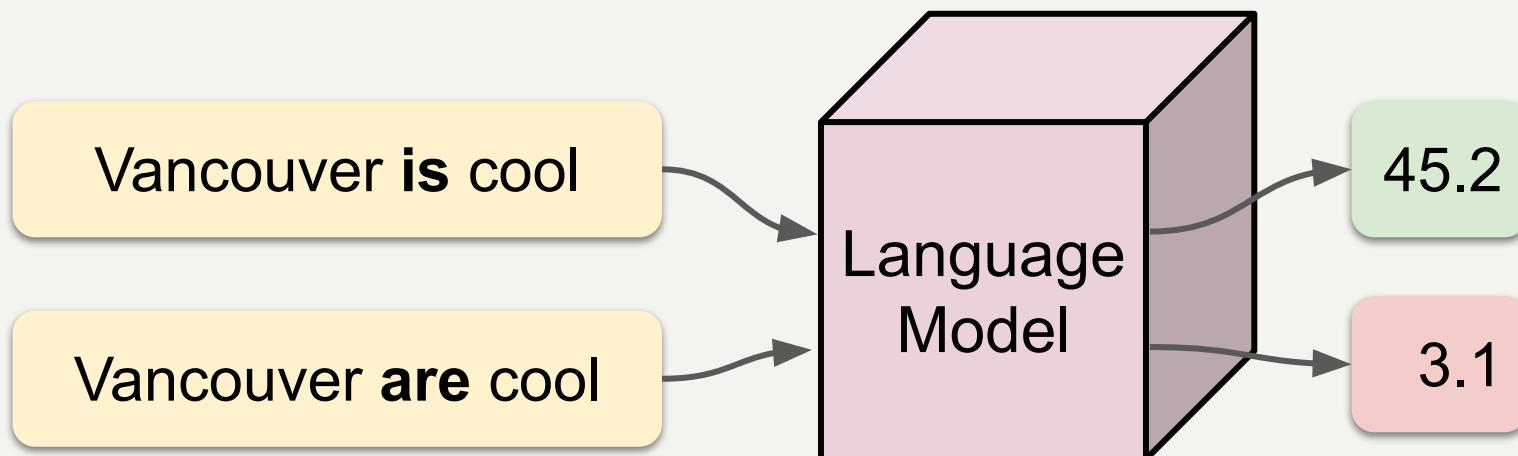
<sup>2</sup>Data Science Group, University of Hamburg, Germany

[vgautam@lsv.uni-saarland.de](mailto:vgautam@lsv.uni-saarland.de)



# Language model

- Our definition: **assigns probabilities** to text
- Lets us distinguish **valid sequences** from invalid ones



## Previous work: Stereotypical biases in language models

Task

The accountant was asked about \_\_\_ charges for preparing tax returns.

$$P(\text{his}) > P(\text{her})$$

# Our work: Measuring pronoun fidelity

## Introduction

The accountant had just eaten a big meal so her stomach was full.

## Task

The accountant was asked about \_\_\_\_\_ charges for preparing tax returns.

$$P(\text{✓ her}) > P(\text{✗ his})$$

# Data construction: 7,200 test items

## 60 occupations

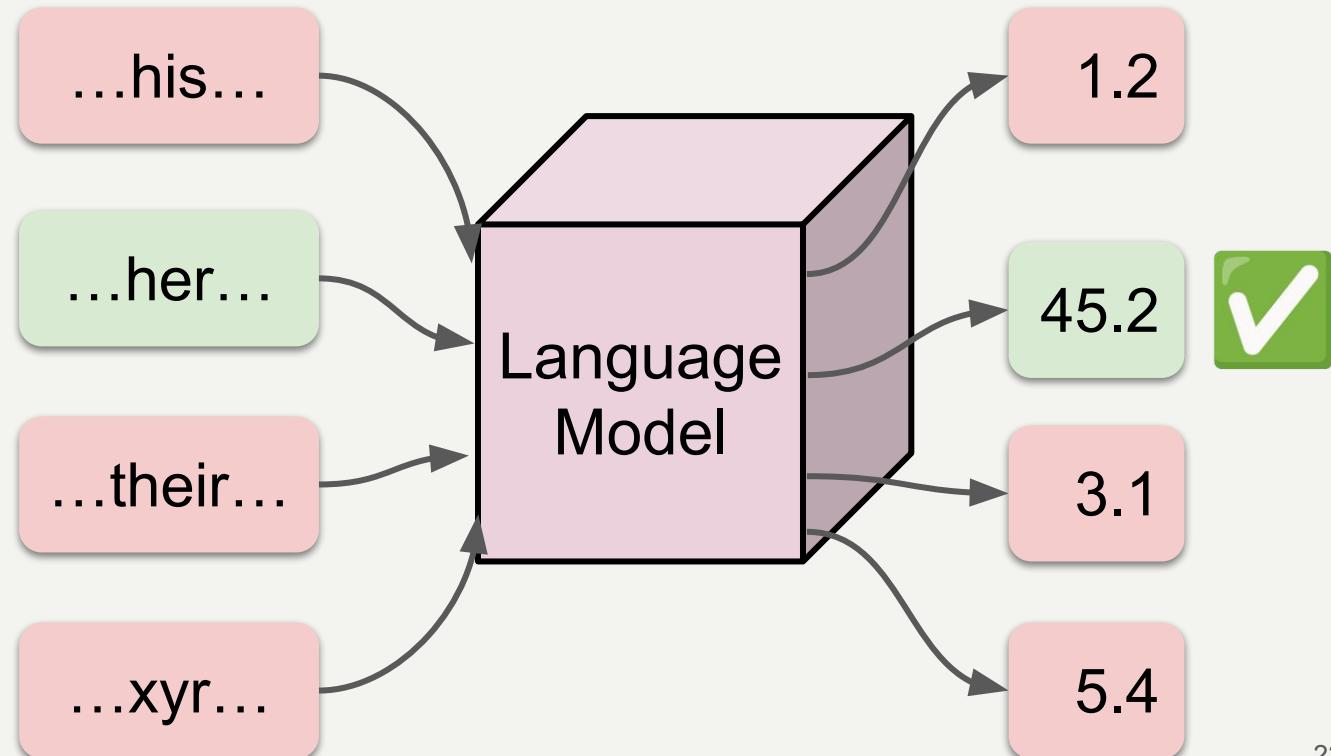
- Chef 
- Scientist 
- Programmer 
- ...

## 3 x 4 pronoun forms

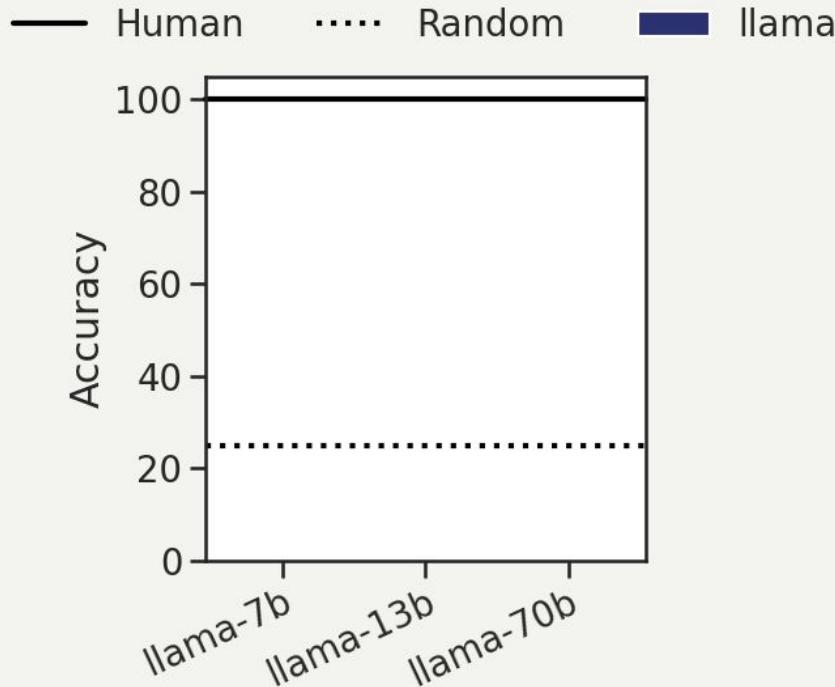
Nominative	Accusative	Possessive
<i>he</i>	<i>him</i>	<i>his</i>
<i>she</i>	<i>her</i>	<i>her</i>
<i>they</i>	<i>them</i>	<i>their</i>
<i>xe</i>	<i>xem</i>	<i>xyr</i>

# Evaluation with language models

The accountant had just eaten a big meal so her stomach was full. The accountant was asked about \_\_\_\_\_ charges for preparing tax returns.

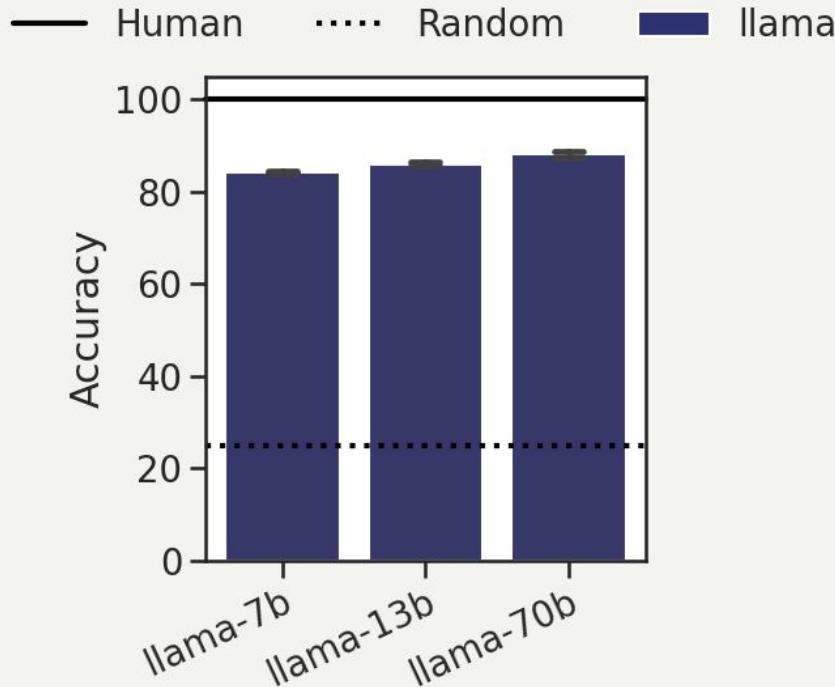


# Results with language models



Compared to random (25%) and human (100%) performance

# Language models show pronoun fidelity!



All models perform above chance (25%) but below humans (100%)

# But is this *true* pronoun fidelity?

- Is this real “**reasoning**” about coreferences?
- Or is it just **repetition** of the pronoun?
- And how can we even **evaluate** this?

But is this *true* pronoun fidelity?

- Is this real “**reasoning**” about coreferences?
- Or is it just **repetition** of the pronoun?
- And how can we even **evaluate** this?

*Add discourse-inspired “distractors”  
and see how models behave!*

# Adding distractors: A second referent!

**Introduction**

The accountant had just eaten a big meal so her stomach was full.

**Distractor(s)**

The **taxpayer** needed coffee because **their** day had started very early.

**Task**

The accountant was asked about \_\_\_\_\_ charges for preparing tax returns.

## Introduction

The accountant had just eaten a big meal so her stomach was full.

## Distractor 1

The taxpayer needed coffee because their day had started very early.

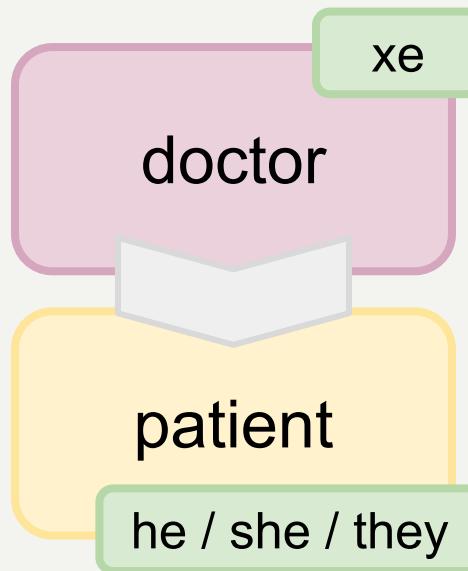
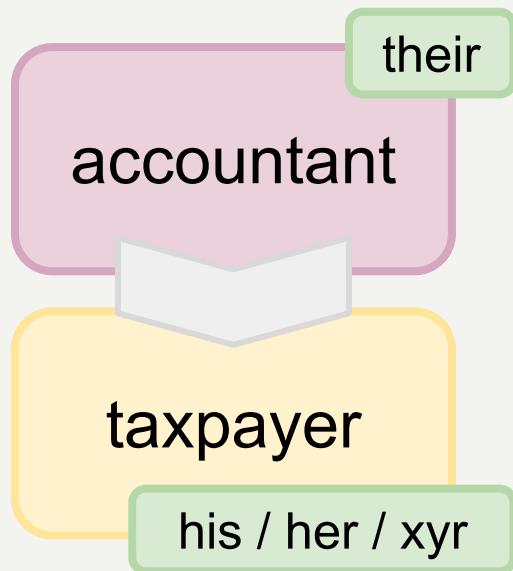
## Distractor 2

Their sleep had been fitful.

## Task

The accountant was asked about \_\_\_\_\_ charges for preparing tax returns.

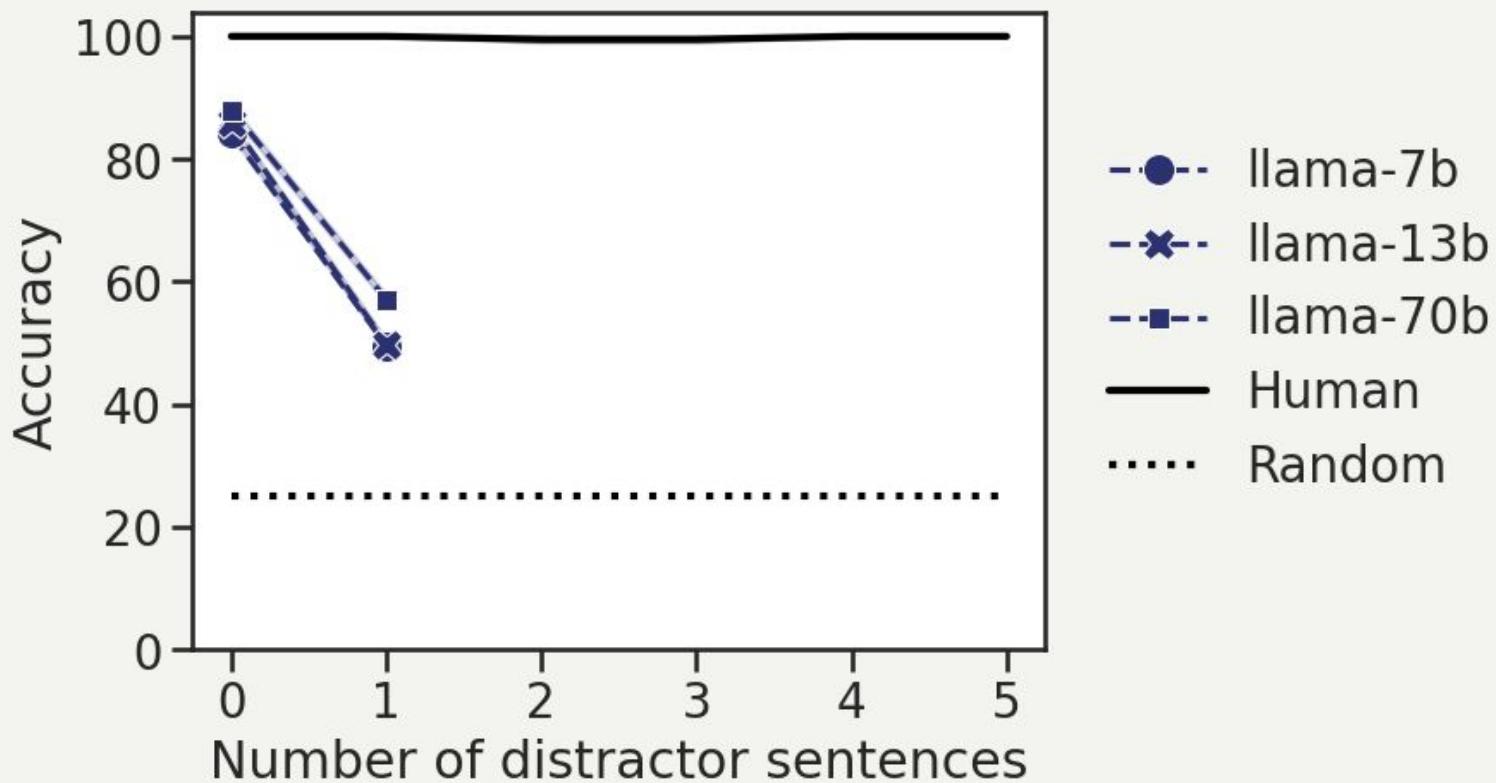
# Data construction: 5 million+ test items



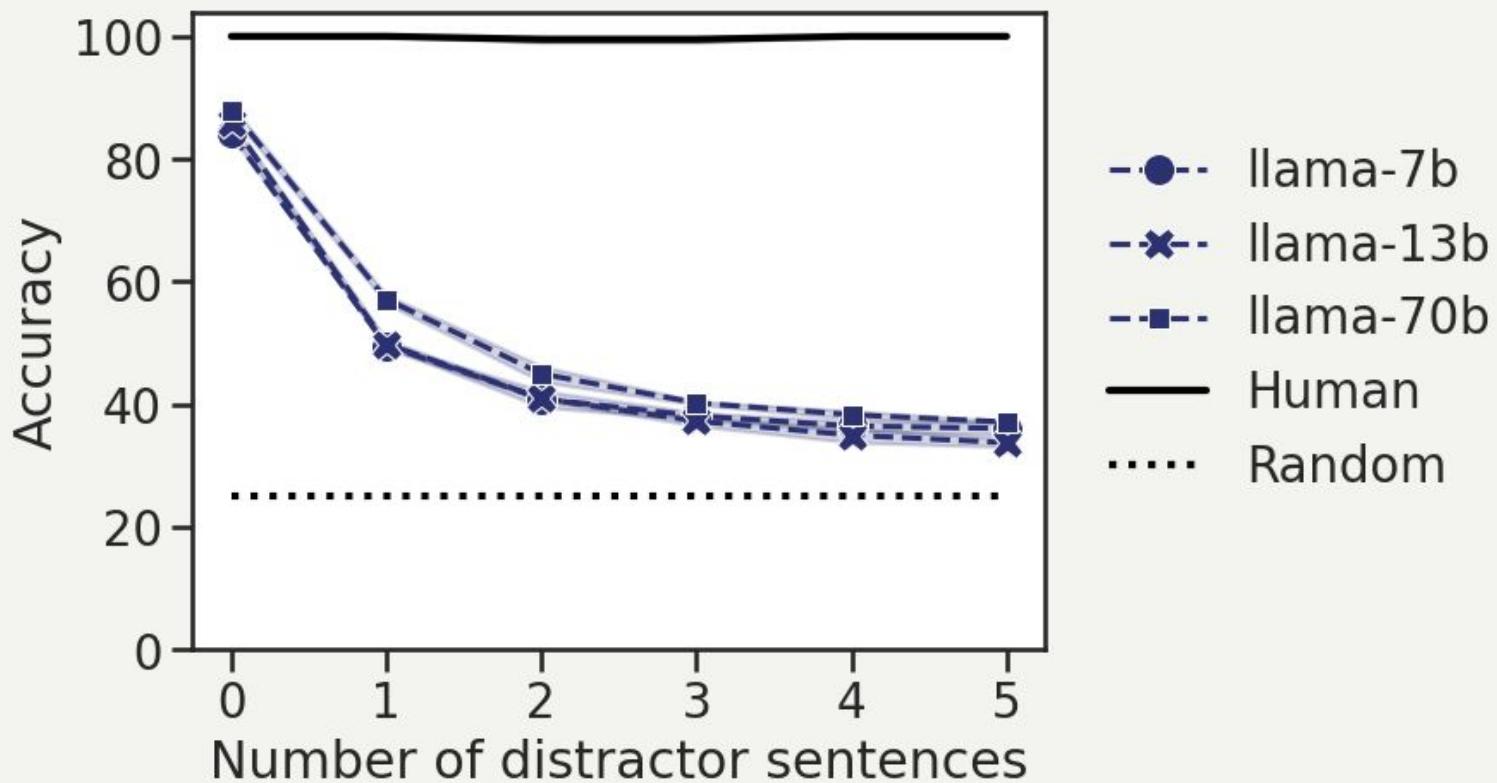
# Pronoun fidelity with distractors



Language models do *not* show pronoun fidelity after all

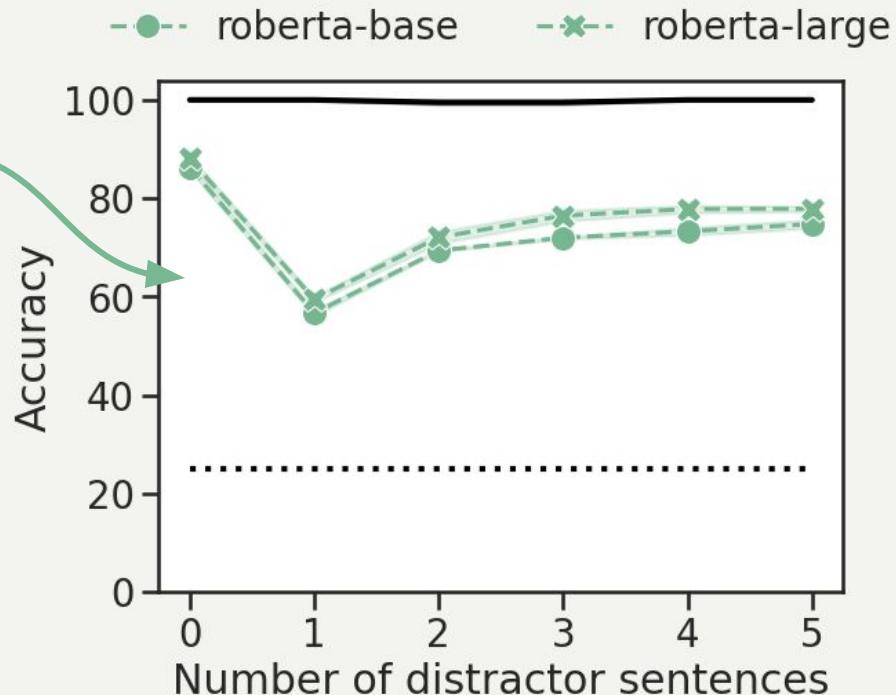


Language models do *not* show pronoun fidelity after all



# Teaser: Bonus material

- 34 more **models** (including **encoder-only**, encoder-decoder)
- Unexpected **scaling** behaviour
- Investigation into the source of **errors**



# Other work: Fairness with referring expressions

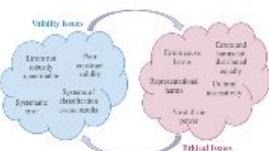
**Stop! In the Name of Flaws:  
Disentangling Personal Names and Sociodemographic Attributes in NLP**

Vagrant Gautam<sup>1</sup> Arjun Subramonian<sup>2</sup> Anne Lauscher<sup>3</sup> Os Keyes<sup>4</sup>  
<sup>1</sup>Saarland University, Germany <sup>2</sup>University of California, Los Angeles, USA  
<sup>3</sup>Universität Hamburg, Germany <sup>4</sup>University of Washington, USA

**Abstract**

Personal names simultaneously differentiate individuals and categorize them in ways that are important in a given society. While the natural language processing community has thus associated personal names with sociodemographic characteristics in a variety of tasks, researchers have engaged to varying degrees with the established methodological problems in doing so. To guide future work that uses names and sociodemographic char-

**Figure 1:** Overview of the methodological issues (cont.)



**Gender Bias in NLP Workshop @ ACL 2024**

CHARACTERISTICS IN NATURAL LANGUAGE PROCESSING

**1 Introduction**

A person's identity is a complex and paradoxical thing – it simultaneously identifies someone's *uniqueness*, and categorizes them, identifying what they have in common with others (Strauss, 2017). A perfect example of this phenomenon is a parent's

In natural language processing (NLP) as well, personal names have a long history of use—NLP researchers have worked on identifying and disambiguating uses of personal names (Mann and Yanowsky, 2003; Minkov et al., 2005; Fürber and Ao, 2022) and have examined name translation (Sennrich et al., 2016; Wang et al., 2022; Kandola et al., 2023) and name fairer resolution (Li

**WINOPRON: Revisiting English Winogender Schemas for Consistency, Coverage, and Grammatical Case**

Vagrant Gautam<sup>1</sup> Julius Steuer<sup>1</sup> Eileen Bingert<sup>1</sup>  
Ray Johns<sup>2</sup> Anne Lauscher<sup>3</sup> Dietrich Klakow<sup>1</sup>  
<sup>1</sup>Saarland University, Germany <sup>2</sup>Independent Researcher, USA  
<sup>3</sup>Data Science Group, University of Hamburg, Germany  
vgautam@cs.uni-saarland.de

**Abstract**

While measuring bias and robustness in coreference resolution are important goals, such measurements are only as good as the tools we use to measure them. Winogender Schemas (Rudinger et al., 2018) are an influential dataset proposed to evaluate gender bias in coreference resolution, but a closer look reveals issues with the data that compromise its

**Winogender Schemas (Rudinger et al., 2018)**

(a) **The counselor disclosed to the patient that he/she had a history of substance abuse.**

(b) **The patient disclosed to the counselor that he/she had a history of substance abuse.**

**WINOPRON (ours)**

(a) **The counselor told the patient that he/she/**they** had a professional duty to report certain issues.**

(b) **The counselor told the patient that he/she/**they** had a professional duty to report certain issues.**

**Computational Models of Reference, Anaphora and Coreference Workshop @ EMNLP 2024**

Third-person pronouns (*he*, *she*, *they*, etc.) help us refer to people in conversation. Since they mark referential gender in English, gender bias affects how coreference resolution systems map these pronouns to people. Rudinger et al. (2018) demonstrated this

evaluation (see Figure 1), which we hypothesize

## Takeaways - *Fair* pronoun usage

- An **evaluation** of **fair** pronoun usage with language models
- Using a discourse-inspired **multi-person setting**
- Language models **do not show pronoun fidelity**, and are easily distracted by other referents
- *We should carefully evaluate claims of “reasoning”*

# Outline

1. Trustworthy NLP?
2. Deep dive: Fairness
3. ↗ **Deep dive: Faithfulness**
4. The future of trustworthy NLP



# Faithfulness?

Broadly, about **encouraging**

- Attention to input
- Grounding in facts and reality
- Following instructions
- Common sense

2 + 3			1
C	%	÷	
7	8	9	×
4	5	6	—
1	2	3	+
0	.		=

Wei et al (2021); Wu et al. (2023); McCoy et al. (2024);  
Rajpurkar et al. (2018); Bastings and Filippova (2020); Jacovi and Goldberg (2020);  
Elazar et al. (2021); Meng et al. (2022); Gupta et al. (2023)

# ChatGPT: What is the capital of Wonderland?

What is the capital of Wonderland?



The capital of Wonderland is often considered to be the town of "Alderlane." It's a whimsical place filled with fantastical characters and curious landscapes, inspired by Lewis Carroll's "Alice's Adventures in Wonderland." If you have a specific version or adaptation in mind, let me know!



# ChatGPT: What is the capital of Wonderland?

No, it's not.

The capita

filled with fantastical characters and curious landscapes, inspired by Lewis Carroll's "Alice's Adventures in Wonderland." If you have a specific version or adaptation in mind, let me know!



How can we teach  
systems to know what  
they don't know?

Findings of EMNLP, 2023

# A Lightweight Method to Generate Unanswerable Questions in English

**Vagrant Gautam**

Saarland Informatics Campus, Saarland University

[vgautam@lsv.uni-saarland.de](mailto:vgautam@lsv.uni-saarland.de)



**Dietrich Klakow**



# What is the capital of British Columbia?

# What is the capital of British Columbia?

The magnifying glass is positioned over the main text body of the Wikipedia article. The text discusses the geographical features, population, and historical significance of British Columbia. It notes its location in the Pacific Northwest, its diverse landscapes, and its status as Canada's third-most populous province.

**British Columbia** [1] is the westernmost province of Canada. Situated in the Pacific Northwest between the Pacific Ocean and the Rocky Mountains, the province has a diverse geography, with rugged landscapes that include rocky coastlines, sandy beaches, forests, lakes, mountains, inland deserts and grassy plains.<sup>[6]</sup> British Columbia borders the province of Alberta to the east; the territories of Yukon and Northwest Territories to the north; the U.S. states of Washington, Idaho, and Montana to the south, and Alaska to the northwest. With an estimated population of over 5.6 million as of 2024, it is Canada's third-most populous province. The capital of British Columbia is Victoria, while the province's largest city is Vancouver. Vancouver and its suburbs together make up the third-largest metropolitan area in Canada, with the 2021 census recording 2.6 million people in Metro Vancouver.<sup>[7]</sup> British Columbia is Canada's third-largest province in terms of total area, after Quebec and Ontario.<sup>[8]</sup>

The first known human inhabitants of the area settled in British Columbia at least 10,000 years ago. Such groups include the Coast Salish, Tsilhqot'in, and Haida peoples, among many others. One of the earliest British settlements in the area was Fort Victoria, established

From Wikipedia, the free encyclopedia  
(Redirected from British Columbia)

Article Talk

Contents hide

(Top)

- > Origin of the name
- > Geography
- > History
- > Demographics
- Economy
- > Government and politics
- > Transportation
- > Arts and culture
- > Outdoor life and athletics
- > Education
- See also
- Notes
- > References
- Further reading
- External links

Search

WIKIPEDIA  
The Free Encyclopedia

Donate Create account Log in ...

Appearance hide

Text

- Small
- Standard
- Large
- Both
- Standard
- Wide

Color (beta)

- Automatic
- Light
- Dark

Motto: Spem Viri Secundum Spem Fecit

Coordinates: 54°N 125°W<sup>[1]</sup>

Canada

Country

BC AB SK MB ON QC PE NL NB NS YT NT NU

# What is the capital of British Columbia?



5.6 million as of 2024, it is Canada's **third-most populous province**.

The capital of British Columbia is **Victoria**, while the province's largest city is **Vancouver**. Vancouver and its suburbs together

# Extractive QA: What is the capital of British Columbia?

Step 1: Document selection

Find the Wikipedia page for  
“British Columbia”

Step 2: Span extraction

Find a sentence about its  
capital

# Extractive QA: What is the capital of British Columbia?

What if you pick the wrong document?

Step 1: Document selection

Find the Wikipedia page for  
“Capital (Marxism)”

Step 2: Span extraction

Find a sentence about its  
???

What if the question has no answer at all?

Extractive QA: What is the capital of **Wonderland**?

Step 1: Document selection

Find the Wikipedia page for  
???

Step 2: Span extraction

Find a sentence about its  
???

In these cases,  
systems typically  
make incorrect  
guesses.

One fix: Build a new system with better data

- **Positive samples:**  
questions that **can be answered** with the given document
- **Negative samples:**  
questions that **cannot be answered** with the document

We get the latter through:

**Unanswerable question generation!**

# Unanswerable question generation

A screenshot of a Wikipedia page for "British Columbia". The page includes a search bar, a sidebar with navigation links like "Contents", "Article talk", and "History", and a main content area with text about the province's origin, geography, and economy. It features images of the provincial flag and coat of arms, and a map of British Columbia's location in Canada. A yellow callout box labeled "Document" is overlaid on the left side of the page.

*What is the capital of  
British Columbia?*

Answerable  
question

The task is to come up with an **unanswerable** question

# Methods to generate unanswerable questions

- **Crowdworkers** (SQuAD<sup>1</sup>)
  - Slow
  - Expensive
  - High quality ✨
- **Language models** (UNANSQ<sup>2</sup>, CRQDA<sup>3</sup>)
  - Fast ✨
  - Computationally intensive
  - Low quality

<sup>1</sup> Rajpurkar et al. (2018) <sup>2</sup> Zhu et al. (2019) <sup>3</sup> Liu et al. (2020)

# Methods to generate unanswerable questions

- **Crowdworkers** (SQuAD<sup>1</sup>)
  - Slow
  - Expensive
  - High quality ✨
- **Language models** (UNANSQ<sup>2</sup>, CRQDA<sup>3</sup>)
  - Fast ✨
  - Computationally intensive
  - Low quality
- RQ: *Can we do this more cleverly?*

<sup>1</sup> Rajpurkar et al. (2018) <sup>2</sup> Zhu et al. (2019) <sup>3</sup> Liu et al. (2020)

# Two ideas from computational linguistics

## 1. Part-of-speech tagging

*Identify the nouns, adjectives, adverbs, etc., in sentences*

## 2. Named entity recognition

*Identify proper nouns in sentences*

Vancouver      is      cool

PROPN      AUX      ADJ

# Antonym swaps

“British Columbia borders the province of Alberta to the east.”

 **Answerable question:**

“What province is east of British Columbia?”

 **Antonym-swapped question:**

“What province is west of British Columbia?”

# Entity swaps

“British Columbia borders the province of Alberta to the east.”

 **Answerable question:**

“What province is east of British Columbia? ”

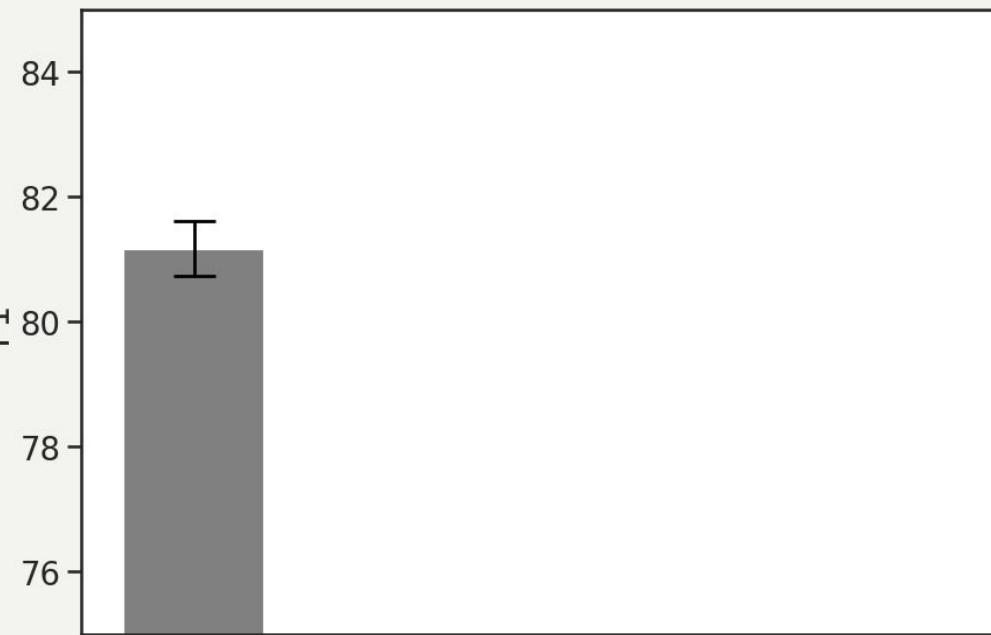
 **Entity-swapped question:**

“What province is east of Alberta? ”

# Can we teach QA systems to say “I don’t know”?

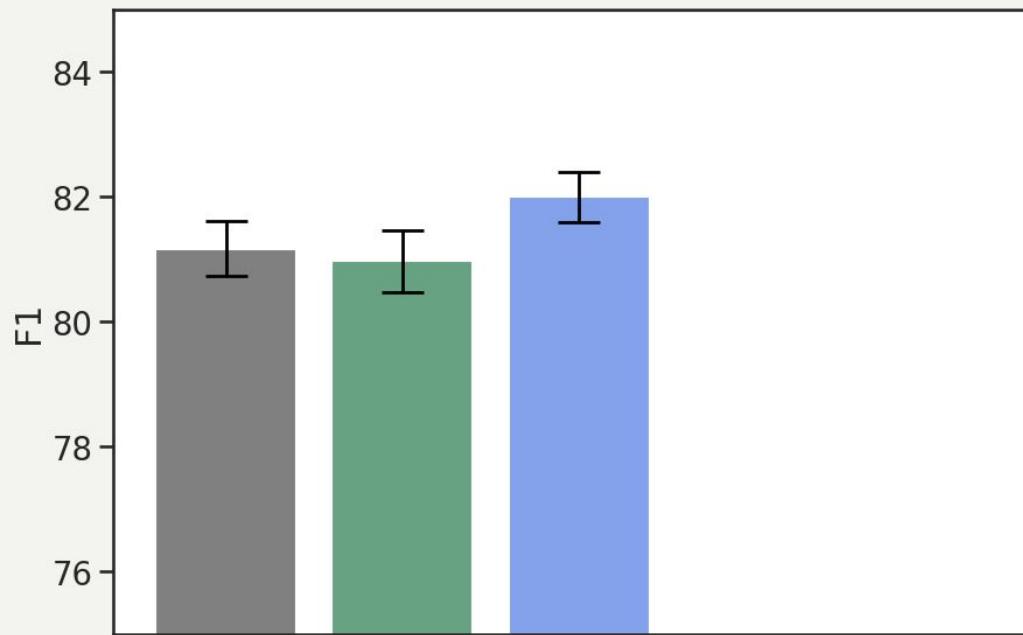
Baseline

Answering  
questions  
*and* saying “I  
don’t know”

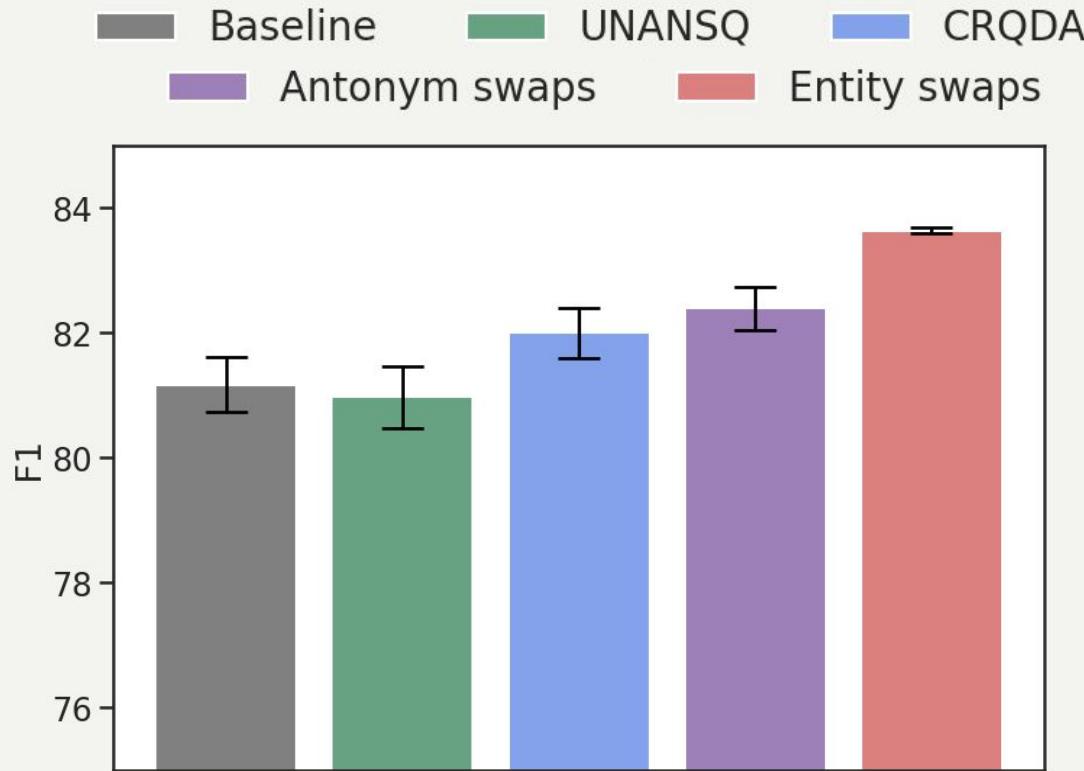


# Language modelling approaches

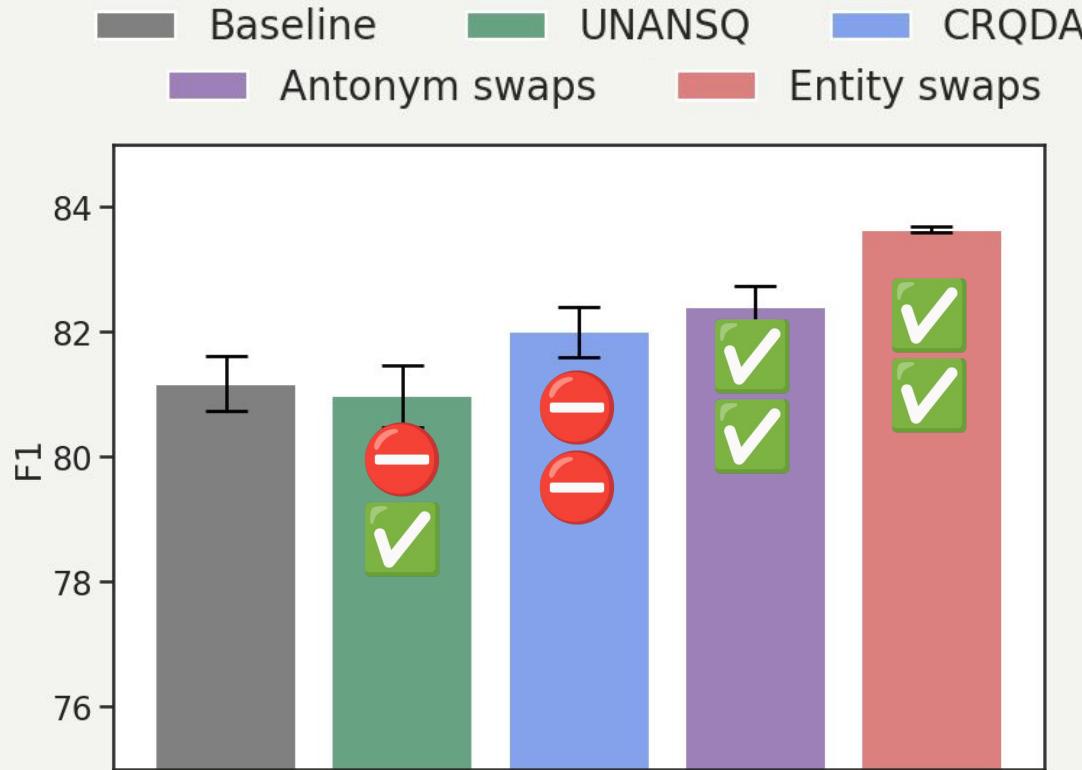
Baseline      UNANSQ      CRQDA



# Our approach is even better!



# Efficiency: Less is more!



## Teaser: Bonus material

“On April 4, 2008, Beyoncé and Jay-Z married without publicity.”

- **Limitations** - where simple swaps would *not* work
- Analysis of the **quality** of generated questions
- More **systems** and **datasets**



## Takeaways - *Faithful* question answering

- A **method** to teach question answering systems to *know what they don't know*
- Using traditional **computational linguistics** techniques
- **Better** in every single way:  
more reliable, more lightweight, better quality data
- *Don't use sledgehammers to crack nuts!*

# Outline

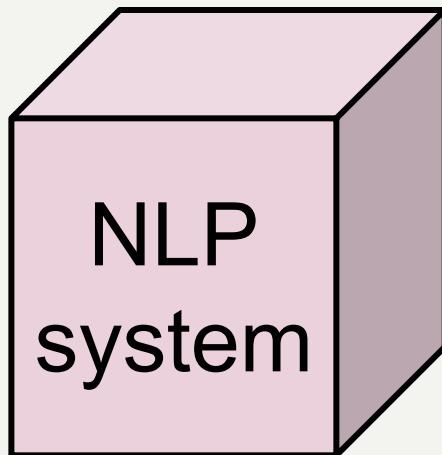
1. Trustworthy NLP?
2. Deep dive: Fairness
3. Deep dive: Faithfulness
4. 🔍 The future of trustworthy NLP



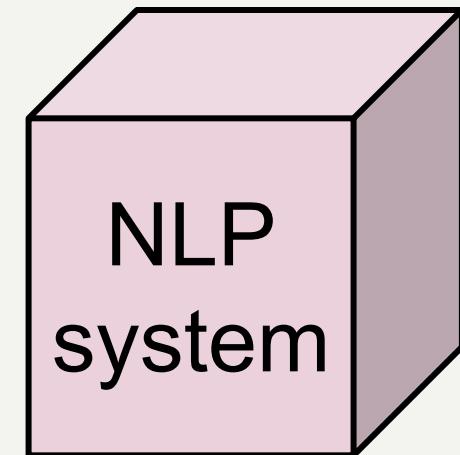
# The future of fairness: Long term

**Human-centered and ethical NLP**

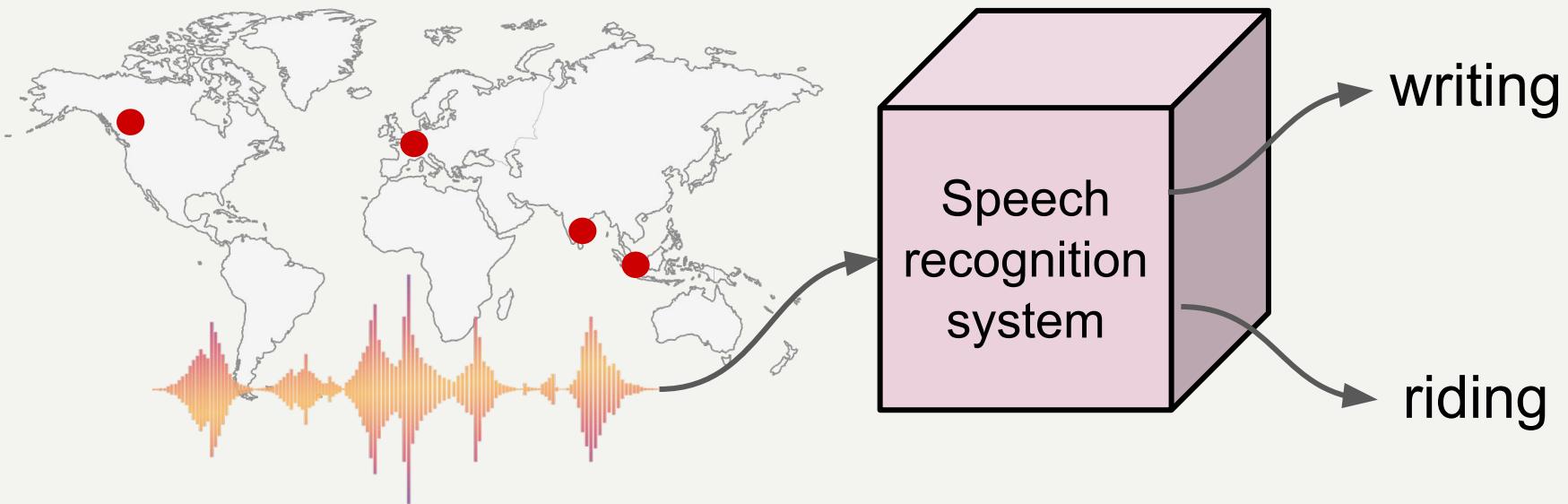
To



or not to

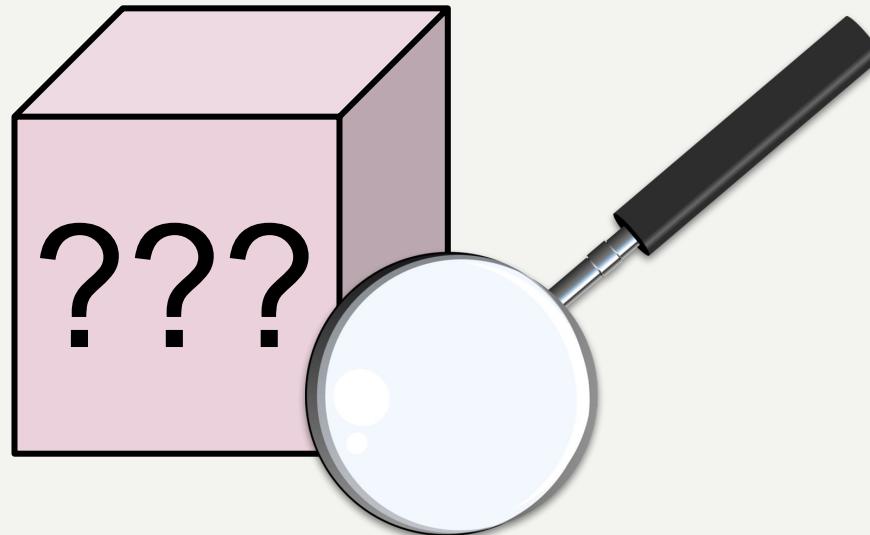


# Variation in speech systems



# The future of faithfulness: Long term

**Interpretable and explainable NLP systems**



# My other work on trustworthy NLP

A word cloud visualization where the size and color of words represent their frequency or importance. The words are arranged in a cluster, with some words having multiple smaller variations nearby.

The words visible in the cloud include:

- syntax
- democratization
- gender
- onomastics
- low-resource
- intersectionality
- survey
- in-context learning
- sociolinguistics
- speech recognition
- history
- grapheme-to-phoneme
- NLP4Science
- interpretability
- phonology
- democracy
- sonority
- names

Thanks to my other PhD collaborators! 🇩🇪 🇨🇭 🇸🇪 🇦🇪 🇺🇸 🇮🇱 🇨🇦 🇲🇽 🇵🇱 🇨🇳



Saarland U

I2SC

DFKI

TU Darmstadt

LMU

HSLU

U of the Basque Country

U of Edinburgh

PUC

MBZUAI

Tel Aviv U

CENIA

EIT Ningbo

McGill U

Mila

UW

UCLA

UC Irvine

Google

Cohere

# Final takeaways

- **Fairness:** Can we trust language models to overcome their biases?
- **Faithfulness:** Teaching question answering systems to know what they don't know

[dippedrusk.com](http://dippedrusk.com)

A word cloud composed of various NLP-related terms in different colors. The words include: syntax, democratization, gender, onomastics, intersectionality, low-resource, survey, in-context learning, sociolinguistics, speech recognition, history, grapheme-to-phoneme, NLP4Science, interpretability, phonology, sonority, democracy, names, and names. The colors used are shades of blue, red, pink, and orange.

We need linguistics for trustworthy NLP systems!

# Backup slides

# Fairness, short term: Pronoun fidelity in German

## Introduction

Der Krankenpfleger hatte Hunger, weil er seit Stunden nichts gegessen hatte.

## Task

Der Krankenpfleger sagte, dass \_\_\_ in einer Stunde Feierabend hätte.

$$P(\text{✓ er}) > P(\text{✗ sie})$$

# Faithfulness, short term: **Meaning-grounded faithfulness**

What is the capital of **Wonderland**?

*Presupposes that:*

>> There exists a capital of Wonderland

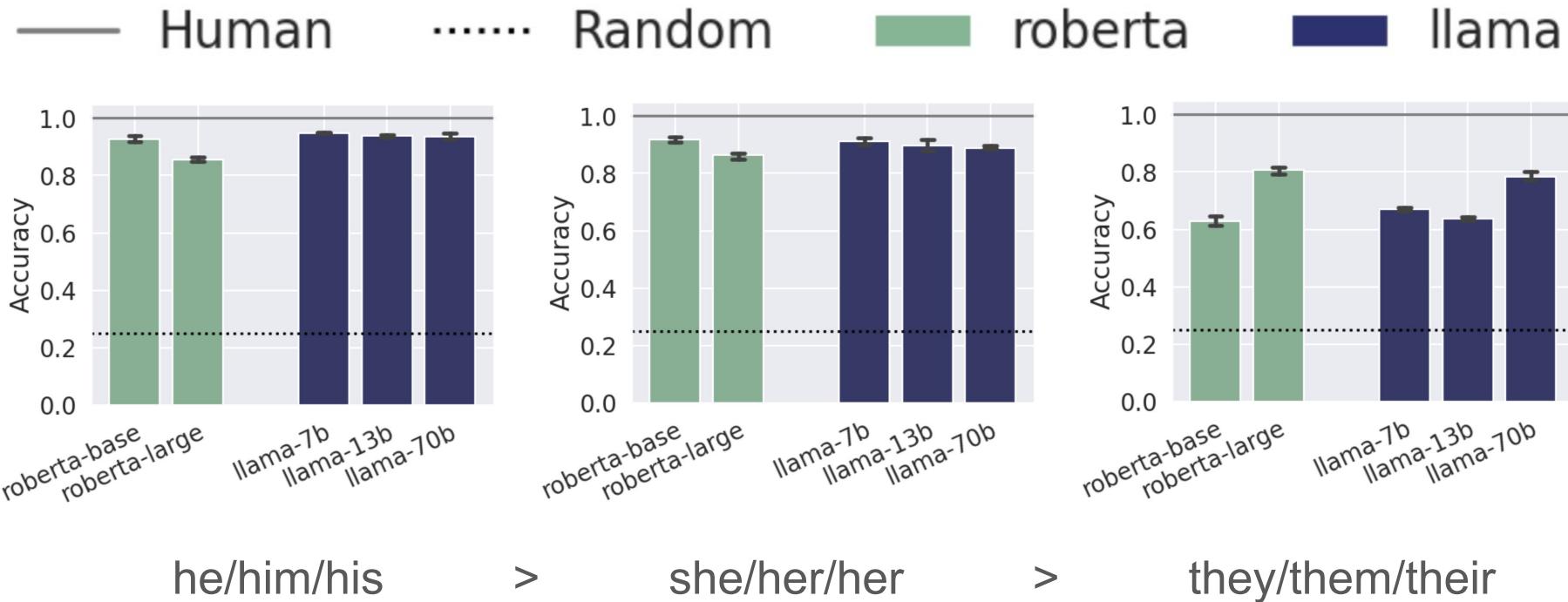
What if we answer the question only if we can  
**verify all of its presuppositions?**<sup>1</sup>

<sup>1</sup> Kim et al. (2021); Kim et al. (2023)

# We evaluate

- Encoder-only models
  - **BERT** - base (110M), large (340M)
  - **ALBERT** - base (**11M**), large (17M), xlarge (58M), xxlarge (223M)
  - **RoBERTa** - base (125M), large (355M)
  - **MosaicBERT** - 137M
- Decoder-only models
  - **OPT** - 125M, 350M, 1.3B, 2.7B, 6.7B, 13B, 30B, 66B
  - **Pythia** - 14M, 70M, 160M, 410M, 1B, 1.4B, 2.8B, 6.9B, 12B
  - **Llama 2** - 7B, 13B, **70B**
- Compared to human (ceiling) and random (baseline) performance

# Models perform best with *he/him/his* pronouns



Are errors due to **bias** or **distraction**?

# Are errors due to bias or distraction?



Model predictions in the absence of context

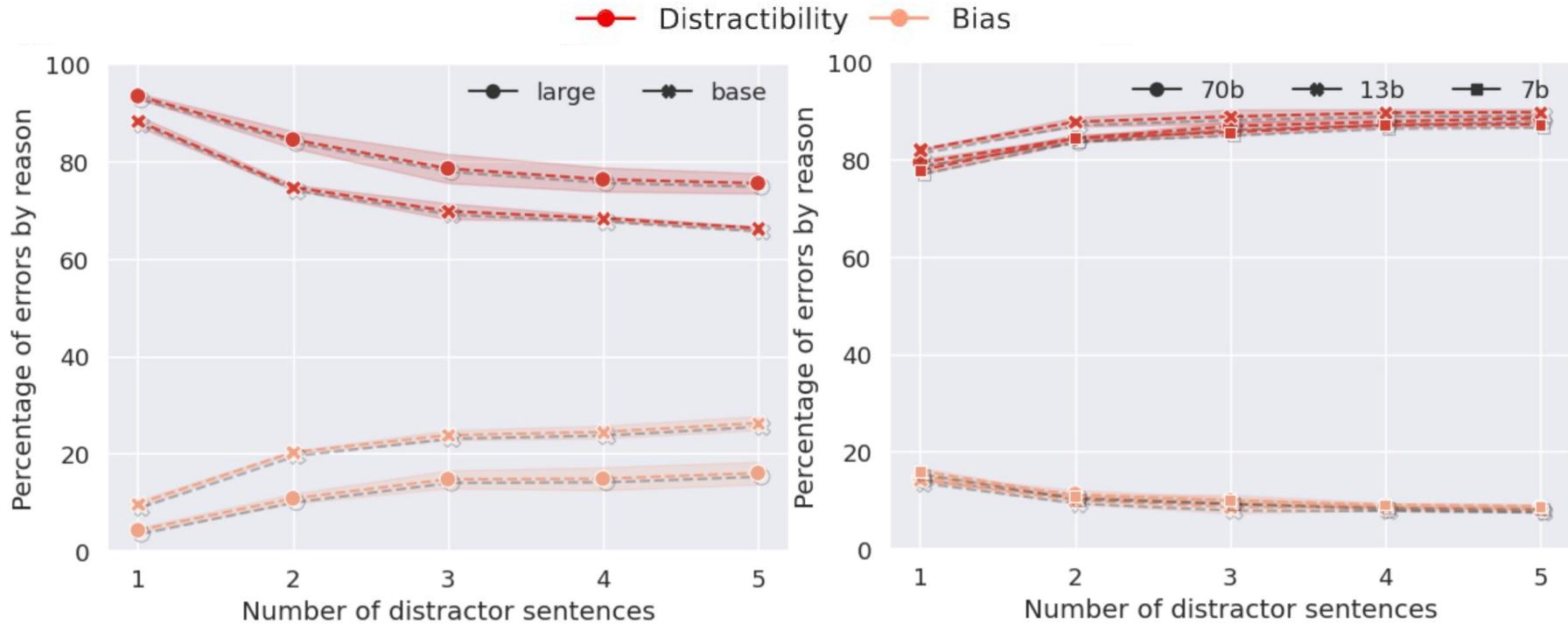
# Are errors due to bias or distraction?

Model predictions in the absence of context

Model predicts the distractor pronoun  
(instead of the biased pronoun or the correct answer)



# Models are mostly distracted, but scale differently



RoBERTa

Llama 2

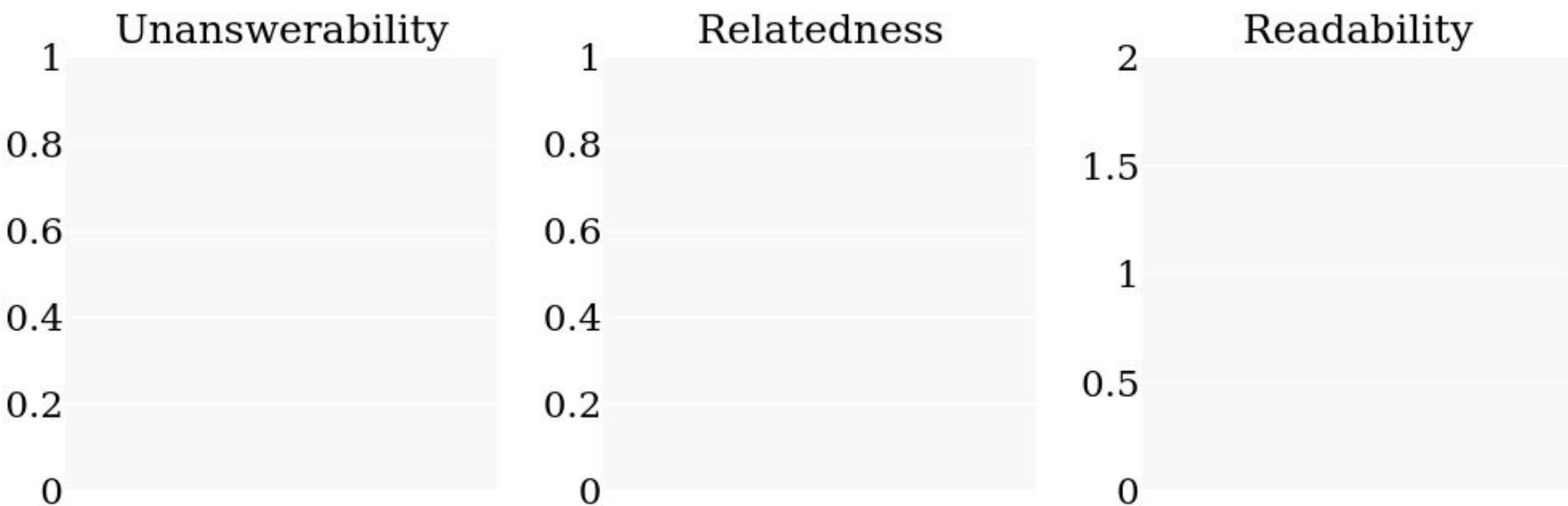
# Limitations

- **Commutative** relations
- **Conjunctions**
- Information is **elsewhere** in the context
- Other forms of **polar questions**
- Dependence on **existing tools**
- Not **extensible** to other languages

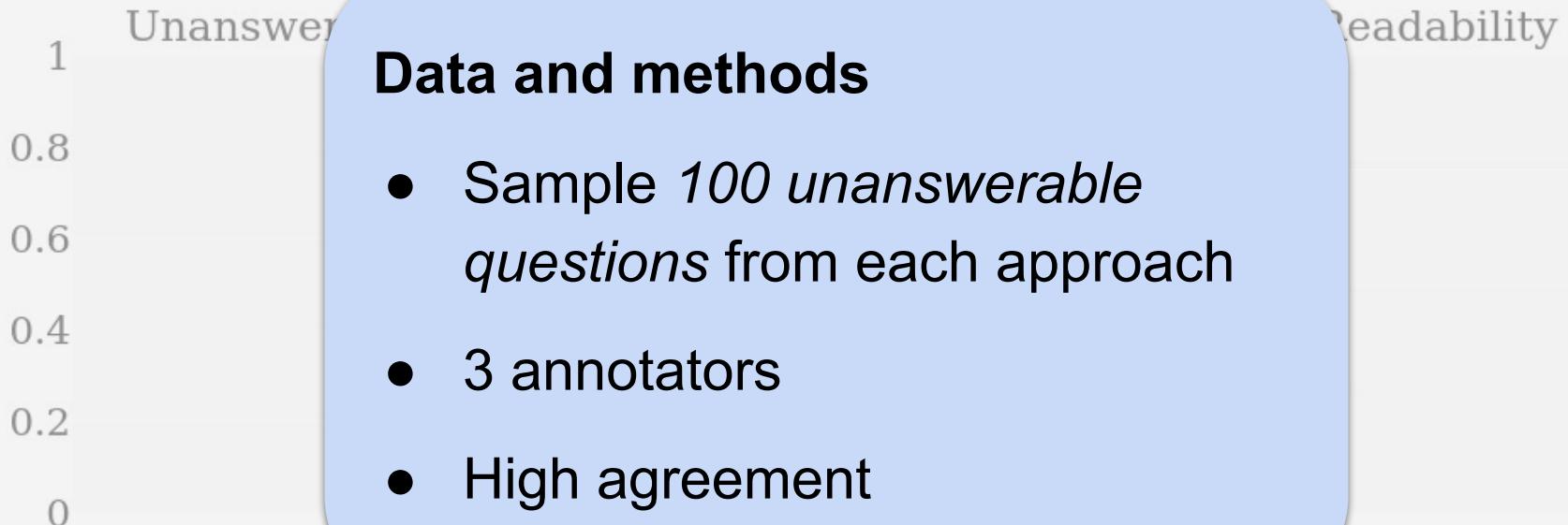
She drives to [Singapore,  
Switzerland, Central Park]

Sie fährt [nach Singapur, in die  
Schweiz, zum Central Park]

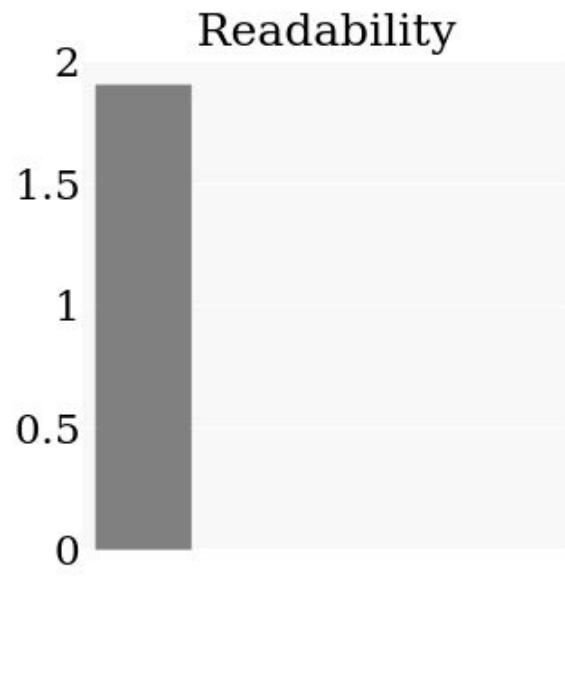
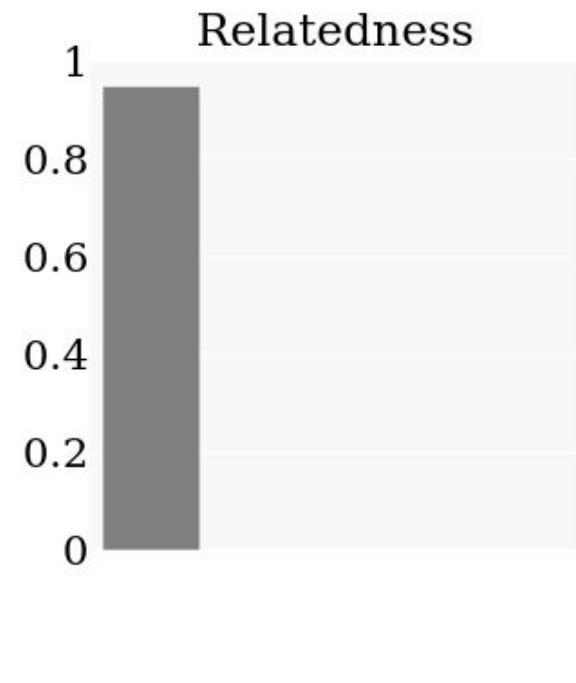
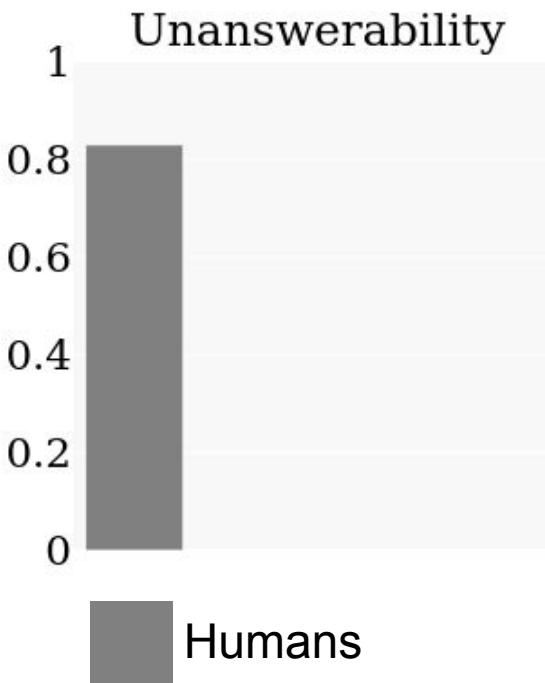
# RQ: How good is our data?



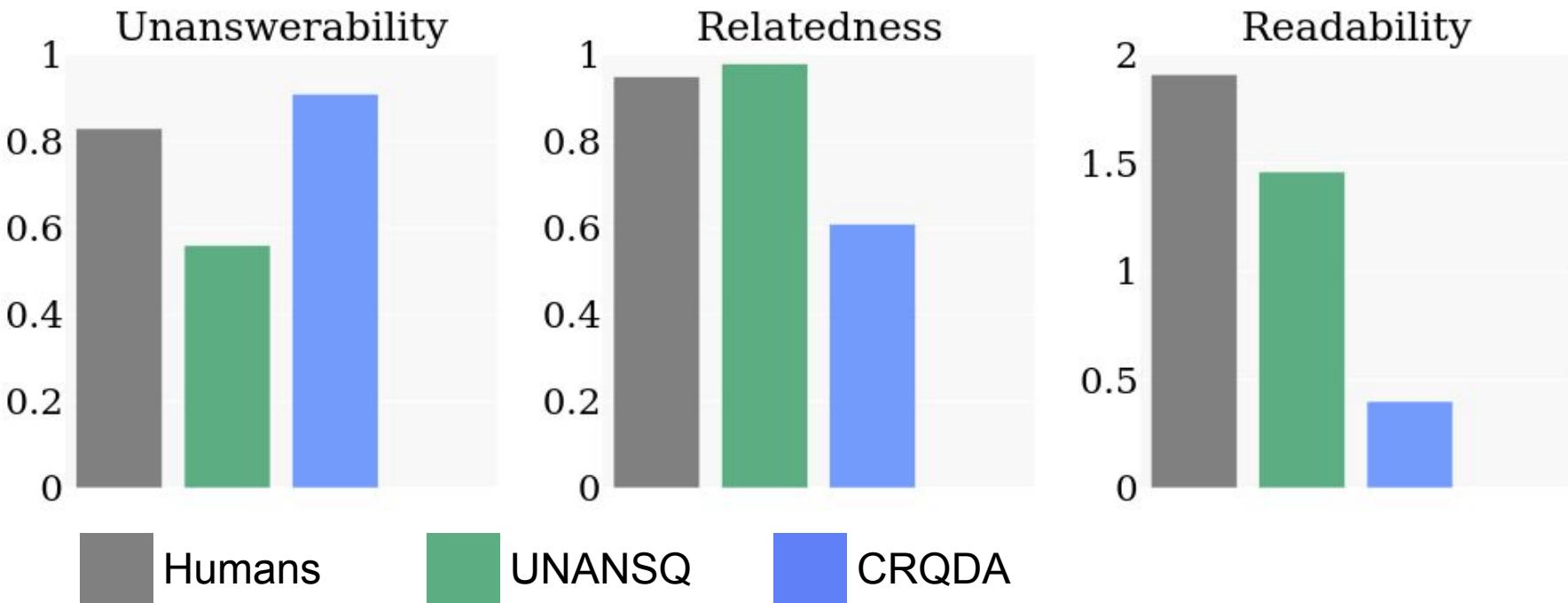
# RQ: How good is our data?



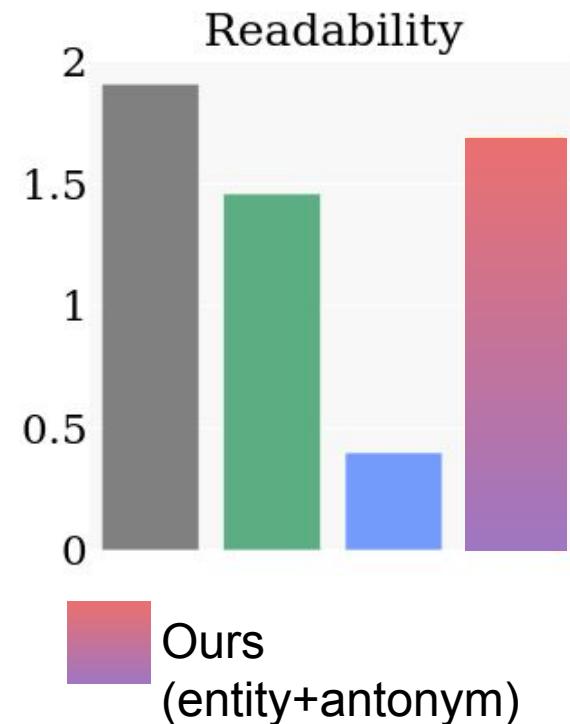
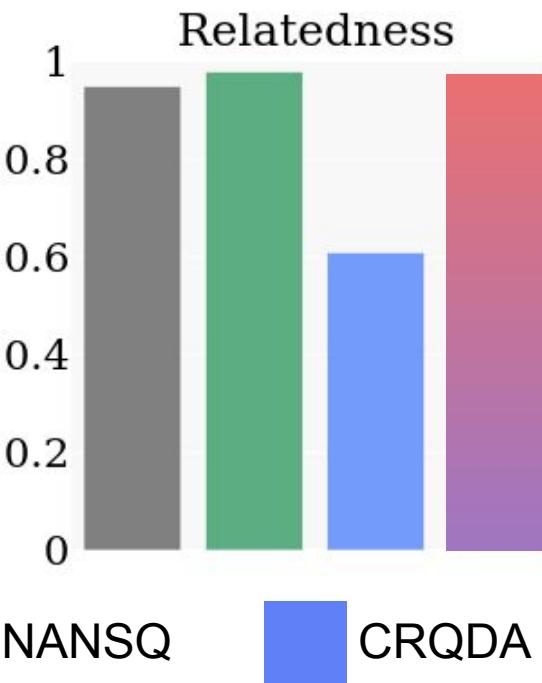
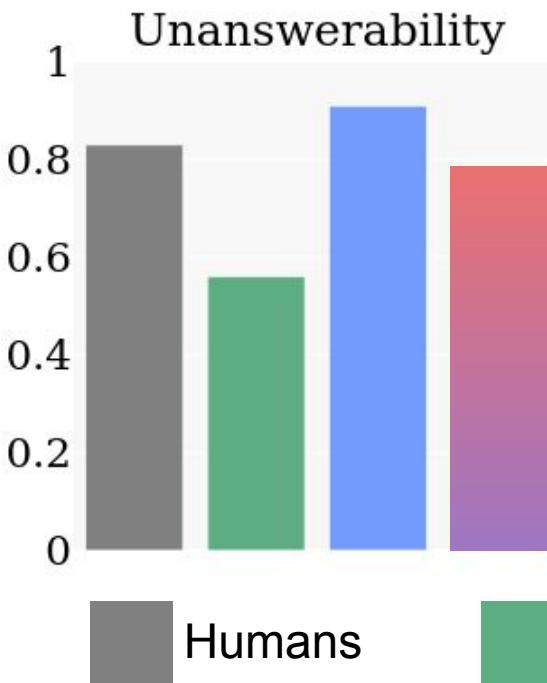
# Humans



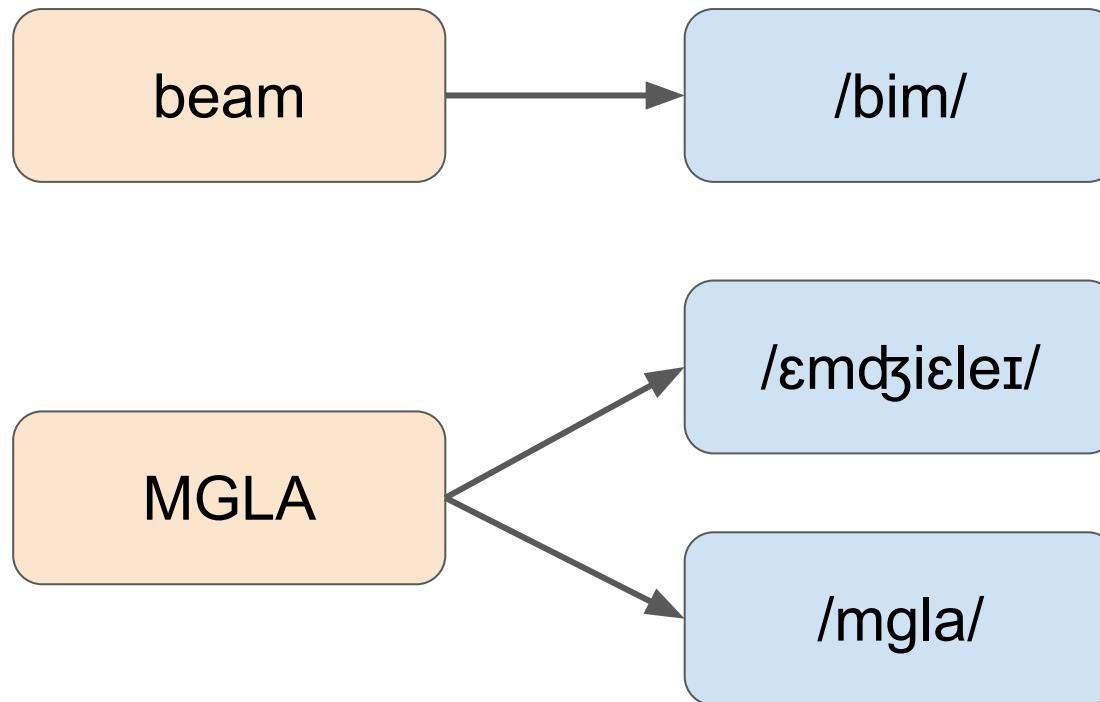
# Language modelling approaches



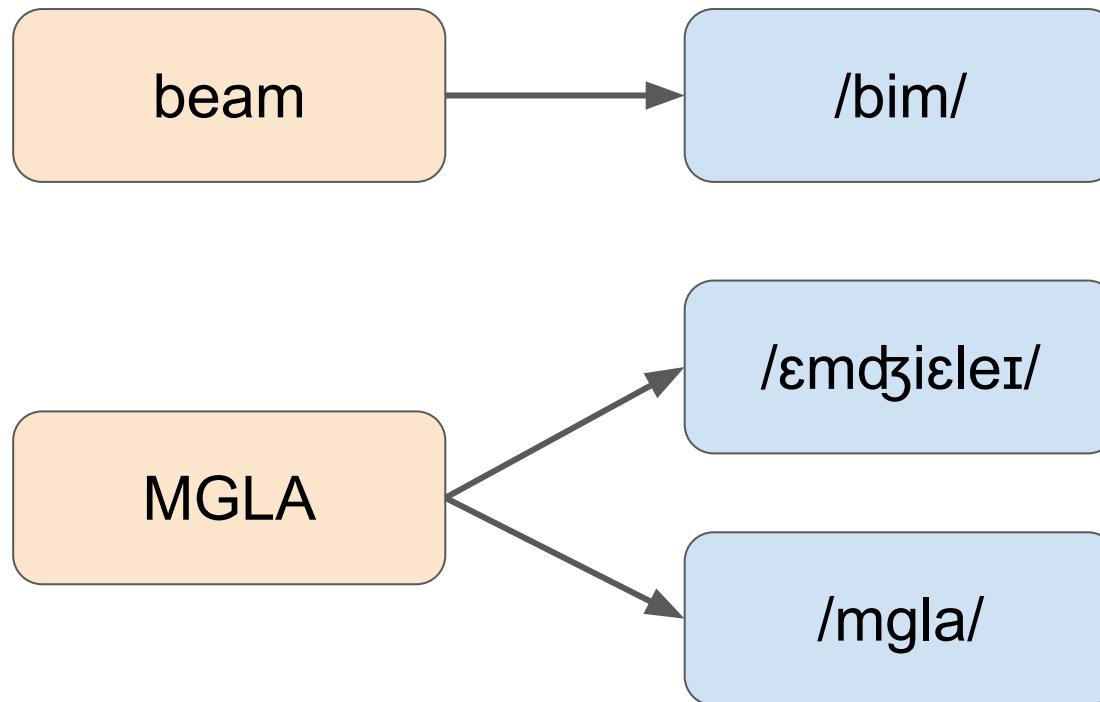
# We're right for the right reasons!



# Using the sonority sequencing principle for G2P



# Using the sonority sequencing principle for G2P



# Coreference resolution

Who does the pronoun *his* refer to in the sentence below?

- (a) The cashier told the customer that his  
card was declined.
- (b) The cashier told the customer that his / her / their  
shift ended soon.

# Coreference resolution

Who does the pronoun *his* refer to in the sentence below?

- (a) The cashier told **the customer** that **his** card was declined.

- (b) The cashier told the customer that his / her / their shift ended soon.

# Coreference resolution

Who does the pronoun *his* refer to in the sentence below?

- (a) The cashier told the customer that his card was declined.
- (b) The cashier told the customer that his shift ended soon.



# Coreference resolution

Who does the pronoun *his* refer to in the sentence below?

- (a) The cashier told the customer that his card was declined.
- (b) The cashier told the customer that his shift ended soon.

# Gender bias in coreference resolution

Vary the pronoun and see how this affects systems!<sup>1</sup>

- (a) The cashier told the customer that his / her / their card was declined.
- (b) The cashier told the customer that his / her / their shift ended soon.

# WinoPron: Findings

- System biases are inconsistent across grammatical case!
- Accusative pronouns are harder to resolve
- Biases are inconsistent across model size

## Winogender Schemas (Rudinger et al., 2018)

(a) The counselor disclosed to the patient that he/she/**they was** professionally mandated to report certain issues.

(b) The patient disclosed to the counselor that he/she/**they** had a history of substance abuse.

## WinoPron (ours)

(1a) The counselor told the patient that he/she/**they/xem** had a professional duty to report certain issues.

(1b) The counselor told the patient that he/she/**they/xem** did not have to pay out of pocket for the sessions.

(2a) The counselor and the patient agreed on having biweekly sessions for him/her/them/**xem** to be able to closely monitor progress.

(2b) The counselor and the patient agreed on having biweekly sessions for him/her/them/**xem** to be able to afford it.

(3a) The counselor informed the patient that his/her/their/**xyr** qualifications were in psychology.

(3b) The counselor informed the patient that his/her/their/**xyr** insurance fully covered the cost of the sessions.

Fix #1: Add 2 missing grammatical cases

Fix #2: Fix structural violations, e.g., non-parallel templates

Fix #3: Ensure templates support all pronouns

Fix #4: Add neopronoun **xem/xyr** to the evaluation

# Personal Names

Personal names are proper nouns used to refer to individuals.

# Personal Names

Personal names are proper nouns used to refer to individuals.

*Individualism* 

- **Uniquely** represent people mentally
- Underscore their significance as **individuals**

# Personal Names: The Paradox

Personal names are proper nouns used to refer to individuals.

*Individualism* 

- **Uniquely** represent people mentally
- Underscore their significance as **individuals**

*Categorization* 

- Socially **organize** communities
- Index societally important **categories**

“Inferences from names must be  
to the givers of these names, not  
to their bearers.”

Weitman (1981)

# Associating names and sociodemographic characteristics is not always a good idea

