



Human-AI Collaboration in Evaluating LLMs

Wei Xu (associate professor)
College of Computing
Georgia Institute of Technology
Twitter/X @cocoweixu



We are obsessed about LLM benchmarks

For very good reasons — collectively, more and better benchmarks and LLMs are coming out!

Meta Llama 3 Instruct model performance

	Meta Llama 3 8B	Gemma 7B - It Measured	Mistral 7B Instruct Measured
MMLU 5-shot	68.4	53.3	58.4
GPQA 0-shot	34.2	21.4	26.3
HumanEval 0-shot	62.2	30.5	36.6
GSM-8K 8-shot, CoT	79.6	30.6	39.9
MATH 4-shot, CoT	30.0	12.2	11.0

	Meta Llama 3 70B	Gemini Pro 1.5 Published	Claude 3 Sonnet Published
MMLU 5-shot	82.0	81.9	79.0
GPQA 0-shot	39.5	41.5 CoT	38.5 CoT
HumanEval 0-shot	81.7	71.9	73.0
GSM-8K 8-shot, CoT	93.0	91.7 11-shot	92.3 0-shot
MATH 4-shot, CoT	50.4	58.5 Minerva prompt	40.5

r/singularity · 19 days ago
Snoo26837

Llama 3 is now top-5 in leaderboard arena.

AI

Rank	Model	Arena Elo	95% CI	Votes	Organization	License
1	GPT-4-Turbo-2024-04-09	1259	+4/-5	23823	OpenAI	Proprietary
1	GPT-4-1106-preview	1254	+3/-3	67933	OpenAI	Proprietary
1	Claude 3 Opus	1252	+3/-3	68656	Anthropic	Proprietary
2	GPT-4-0125-preview	1249	+3/-3	56475	OpenAI	Proprietary
5	Meta Llama 3 70b Instruct	1210	+5/-5	12719	Meta	Llama 3 Community
5	Bard (Gemini Pro)	1208	+6/-6	12435	Google	Proprietary
5	Claude 3 Sonnet	1202	+2/-3	70952	Anthropic	Proprietary
8	Command R+	1192	+3/-4	39243	Cohere	CC-BY-NC-4.0
8	GPT-4-0314	1189	+3/-3	46299	OpenAI	Proprietary
10	Claude 3 Haiku	1181	+3/-3	64106	Anthropic	Proprietary
11	GPT-4-0613	1165	+3/-3	65048	OpenAI	Proprietary
12	Mistral-Large-2402	1158	+3/-3	42206	Mistral	Proprietary

Meta Llama 3 is now top-5 in Arena!

422 ↓ 122 Share

Can we do better to evaluate & create LLMs?

Goal 1 - User Satisfaction



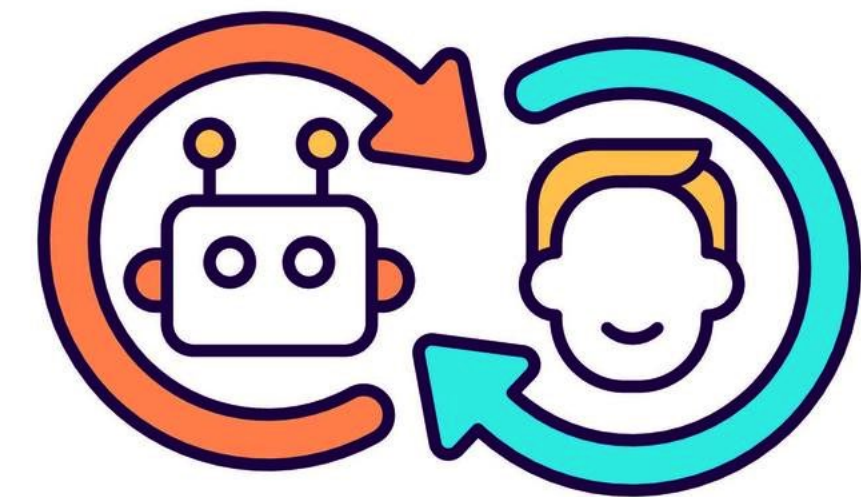
Conduct interview-based user study to solicit feedback that informs AI design

Goal 2 - Global Equity



Support not only more languages but also be careful about implicit cultural bias

Goal 3 - Interactive Interface



Design user interface to support more sophisticated human evaluation

Today's talk — three case studies

Goal 1 - User Satisfaction

PrivacyMirror



(Yao et al., 2024)

Conduct interview-based user study to solicit feedback that informs AI design

Goal 2 - Global Equity

CAMEL



(Naous et al., 2024)

Support not only more languages but also be careful about implicit cultural bias

Goal 3 - Interactive Interface

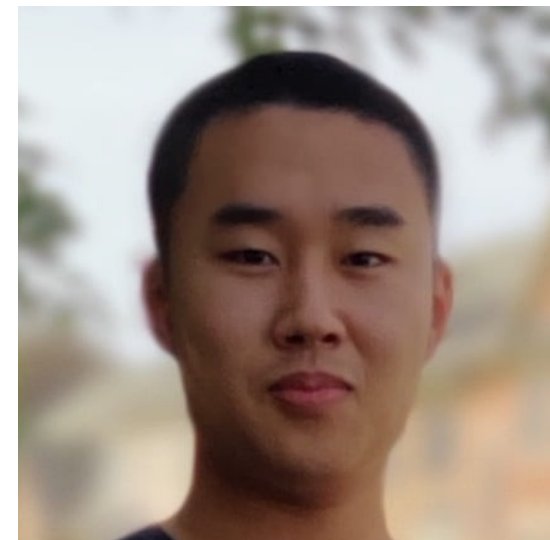
THRESH



(Heineman et al., 2023)

Design user interface to support more sophisticated human evaluation

Reducing Privacy Risks in Online Self-Disclosures (PrivacyMirror 🪞)



Yao Dou



Isadora Krsek



Tarek Naous



Anubha Kabra



Sauvik Das



Alan Ritter



Wei Xu

People talk about themselves online

↑ Posted by u/[deleted] 7 months ago

19 **For those who joined the military to find your way, where are you now?**

↓ Advice

 KnightCPA · 7 mo. ago

I joined at 23. I'm now a DV. I had a good career, over 13 years as a medic. There's a lot to unpack, but it can be either a good career or a valuable stepping stone, or launch point. It can also cause problems if you are undisciplined. My only regret is not having an understanding of the pipelines that interested me the most when I joined. I didn't quite do everything I wanted to do before my time was over. Before going in, start planning. Which branches interest you? Next what kind of jobs interest you? Perhaps the most important is, what obligations could potentially hold you back. Are you divorced with 3 kids from multiple partners? Do you have any critical vices? Are you a felon? Take care of any of these issues before you go, that way you can focus on training.

You will earn 30 days of vacation per year, a bonus for joining (potentially), a steady pay check, \$4500/yr tuition assistance and more opportunities than you will be able to take advantage of. However, you will deal with power tripping ego-maniacs, orders based on political whims, and questionable ethics regularly.

I was fortunate to have the oppositely to travel the world, a couple of times. For me it was worth it. In fact, I should have joined sooner. I am now two years out of service and seeking a new career. This last part is the last great challenge, so far as I can tell, for my future. For me, I would do it again, and I would do it differently. However, I hope to provide my son every opportunity to keep him from feeling obligated, or influenced to serve. I want to make one thing very clear: military service is NOT a typical 9-5, 40hr/week job. Feel free to DM me with any questions.

↑ 2 ↓  Reply  Share ...

People talk about themselves online

↑ Posted by u/[deleted] 7 months ago

19 For those who joined the military to find your way, where are you now?

↓ Advice

 KnightCPA · 7 mo. ago

I joined at 23. I'm now a DV. I had a good career, over 13 years as a medic. There's a lot to unpack, but it can be either a good career or a valuable stepping stone, or launch point. It can also cause problems if you are undisciplined. My only regret is not having an understanding of the pipelines that interested me the most when I joined. I didn't quite do everything I wanted to do before my time was over. Before going in, start planning. Which branches interest you? Next what kind of jobs interest you? Perhaps the most important is, what obligations could potentially hold you back. Are you divorced with 3 kids from multiple partners? Do you have any critical vices? Are you a felon? Take care of any of these issues before you go, that way you can focus on training.

You will earn 30 days of vacation per year, a bonus for joining (potentially), a steady pay check, \$4500/yr tuition assistance and more opportunities than you will be able to take advantage of. However, you will deal with power tripping ego-maniacs, orders based on political whims, and questionable ethics regularly.

I was fortunate to have the opportunity to travel the world, a couple of times. For me it was worth it. In fact, I should have joined sooner. I am now two years out of service and seeking a new career. This last part is the last great challenge, so far as I can tell, for my future. For me, I would do it again, and I would do it differently. However, I hope to provide my son every opportunity to keep him from feeling obligated, or influenced to serve. I want to make one thing very clear: military service is NOT a typical 9-5, 40hr/week job. Feel free to DM me with any questions.

↑ 2 ↓ Reply Share ...

Disclosures:

1. Join army at 23
2. Now a DV (distinguished visitor)
3. Over 13 years as a medic
4. No job, out of service 2 years
5. Has a son

Prior Work on Privacy Preservation

PII Identification and Anonymization ([Lukas et al. 2023](#), [Lison et al. 2021](#), and more)

- Highly-sensitive personal information that are common in medical or legal texts

ACCOUNT TRANSFER REQUEST

To,
The Branch Manager
20520
Bank of America

From: Name: **Mustafa Abdul**
Address: **2201 C Street NW I Washinton, DC**
Phone No.: **797-861-7797**

Madam/ Dear Sir,

Request for my /our SB/RD/Term Deposit Account Transfer
A/c No. **GL28 0219 2024 5014 48**
From (Branch Name- Code) to (Branch Name- Code)

1. I hold the above account/accounts with branch code: **BOFAUS3N.**
2. I request you to transfer the captioned account. The new address proof is enclosed/ shall be provided within 6 months at the transferee branch.
3. I request you to transfer the CIF.
4. I understand that if CIF is not transferred, my Home Branch will continue to remain the same.

Please arrange accordingly.

Yours faithfully,

Mustafa Abdul
Dated: 11th Jan, 2018

- Personal**
 - Full name
 - Home address
 - Face
 - Phone number
 - Date of birth
 - Email
 - First name
 - Last name
 - Street
 - City
 - Country
- Health**
 - Personal health information (PHI)
 - Medical records
 - WHO ICD codes
- National**
 - Passport
 - Driving license
 - SSN
 - Tax ID
- Financial**
 - Bank account number
 - Credit card number
 - Routing number
- Security**
 - Username
 - Password
 - IP address
- Sensitive**
 - Sexual preferences
 - Political views
 - Race
 - Gender
 - Religious view
- Custom**
 - Define your own detection patterns

- Existing tools often detect “non-personal” information indiscriminately

“Freelance illustrator taking commissions. Contact me at xxxxyyyzzz@gmail.com”

PrivacyMirror — provide user-side protection

Detection

Abstraction

PrivacyMirror — provide user-side protection

Detection

I joined at 23. I'm now a Distinguished Visitor. I had a good career, over 13 years as a medic.

Abstraction

PrivacyMirror — provide user-side protection

Detection

I joined at 23. I'm now a Distinguished Visitor. I had a good career, over 13 years as a medic.

Abstraction

PrivacyMirror — provide user-side protection

Detection

I joined at 23. I'm now a Distinguished Visitor. I had a good career, over 13 years as a medic.

Abstraction

I joined at 23. → *I joined when I was in my early twenties.*

I'm now a Distinguished Visitor.

→ *I'm currently holding a prestigious title.*

over 13 years as a medic.

→ *have spent many years in the medical field*

PrivacyMirror — 19 Self-disclosure Categories

We manually annotated and categorized 4.8K annotated self-disclosures that are beyond PII.

Demographic Attributes

Age

Age&Gender

Race/Nationality

Gender

Location

Appearance

Wife/GF

Husband/BF

Sexual Orientation

Relationship Status

Pet

Contact

Name

Personal Experiences

Occupation

Family

Health

Mental Health

Finance

Education

PrivacyMirror — 19 Self-disclosure Categories

We manually annotated and categorized 4.8K annotated self-disclosures that are beyond PII.

I live in the UK and a diagnosis is really expensive,...

Same here. I am 6'2. No one can sit behind me.

I'm a straight man but I do wanna say this

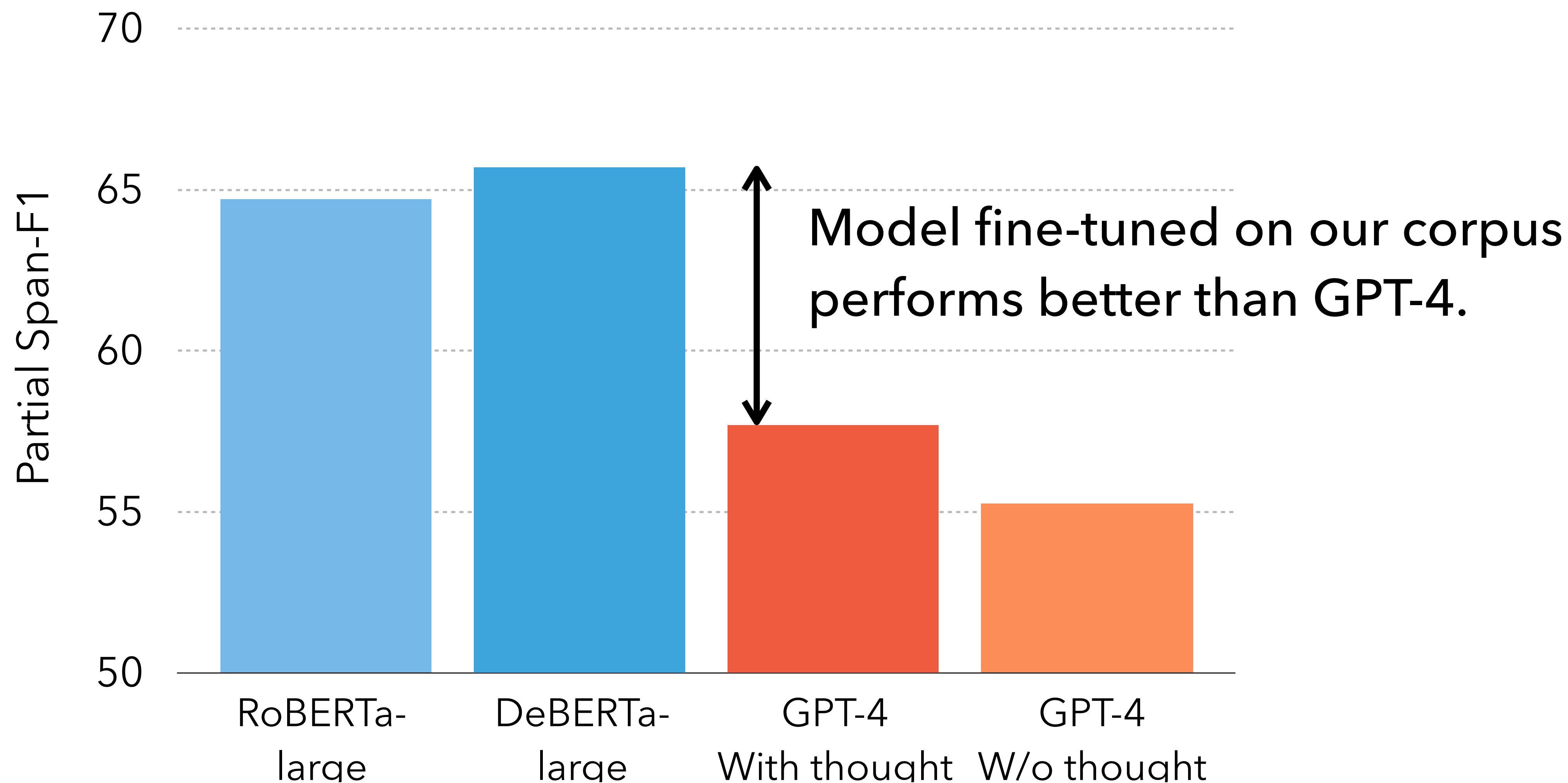
Hi there, I got accepted to UCLA (IS), which I'm pumped about.

My little brother (9M) is my pride and joy

My husband and I vote for different parties

PrivacyMirror — Self-disclosure Detection

We can train automatic detection models by fine-tuning on our corpus or prompting GPT-4.



PrivacyMirror — Do real users like our tool?

We interviewed 21 Reddit users for ~2 hours. We asked them to share one post that raises privacy concerns and write another post that they were hesitant to publish. Then we run our model.

82% participants view the model **positively**

Interesting Feedback

Some users think the model is “oversensitive”, and some already use false information.

→ Personalization and Rate Importance

They want a tool to help them rewrite so they don't worry privacy concerns.

→ Abstraction

PrivacyMirror — Self-disclosure Abstraction

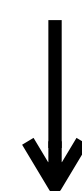
Rephrases disclosures with less specific details while preserving the content utility.

Sentence: Not 21 so can't even drink really even tho I'm in Korea.

PrivacyMirror — Self-disclosure Abstraction

Rephrases disclosures with less specific details while preserving the content utility.

Sentence: Not 21 so can't even drink really even tho I'm in Korea.



Not of legal drinking age



I'm abroad.

PrivacyMirror — Self-disclosure Abstraction

Rephrases disclosures with less specific details while preserving the content utility.

Sentence: Not 21 so can't even drink really even tho I'm in Korea.



Not of legal drinking age



I'm abroad.

Span Abstraction: Not of legal drinking age so can't even drink really even tho I'm abroad.

PrivacyMirror — Self-disclosure Abstraction

Comparing span-level “abstraction” to other sentence-level “abstraction” methods.

Sentence: Not 21 so can't even drink really even tho I'm in Korea.

Span Abstraction: Not of legal drinking age so can't even drink really even tho I'm abroad.

PrivacyMirror — Self-disclosure Abstraction

Comparing span-level “abstraction” to other sentence-level “abstraction” methods.

Sentence: Not 21 so can't even drink really even tho I'm in Korea.

Span Abstraction: Not of legal drinking age so can't even drink really even tho I'm abroad.

Anonymization: [xxx] so can't even drink really even tho [xxx]

Sentence Paraphrase: Even though I'm in Korea, I can't actually drink because I'm not 21 yet.

Sentence Abstraction: Not old enough to legally consume alcohol even though I'm abroad.

PrivacyMirror — Self-disclosure Abstraction

Comparing span-level “abstraction” to other sentence-level “abstraction” methods.

Sentence: Not 21 so can't even drink really even tho I'm in Korea.

Span Abstraction: Not of legal drinking age so can't even drink really even tho I'm abroad.

Anonymization: [xxx] so can't even drink really even tho [xxx] ~~Utility~~

Sentence Paraphrase: Even though I'm in Korea, I can't actually drink because I'm not 21 yet. ~~Privacy~~

Sentence Abstraction: Not old enough to legally consume alcohol even though I'm abroad.

~~Writing Style~~

PrivacyMirror — Self-disclosure Abstraction

Comparing span-level “abstraction” to other sentence-level “abstraction” methods.

Sentence: Not 21 so can't even drink really even tho I'm in Korea.

Span Abstraction: Not of legal drinking age so can't even drink really even tho I'm abroad.

✓ Utility

✓ Privacy

✓ Writing Style

[xx] so can't even drink really even tho [xxx] Utility

Sentence Par

h I'm in Korea, I can't actually drink because I'm not 21 yet. Privacy

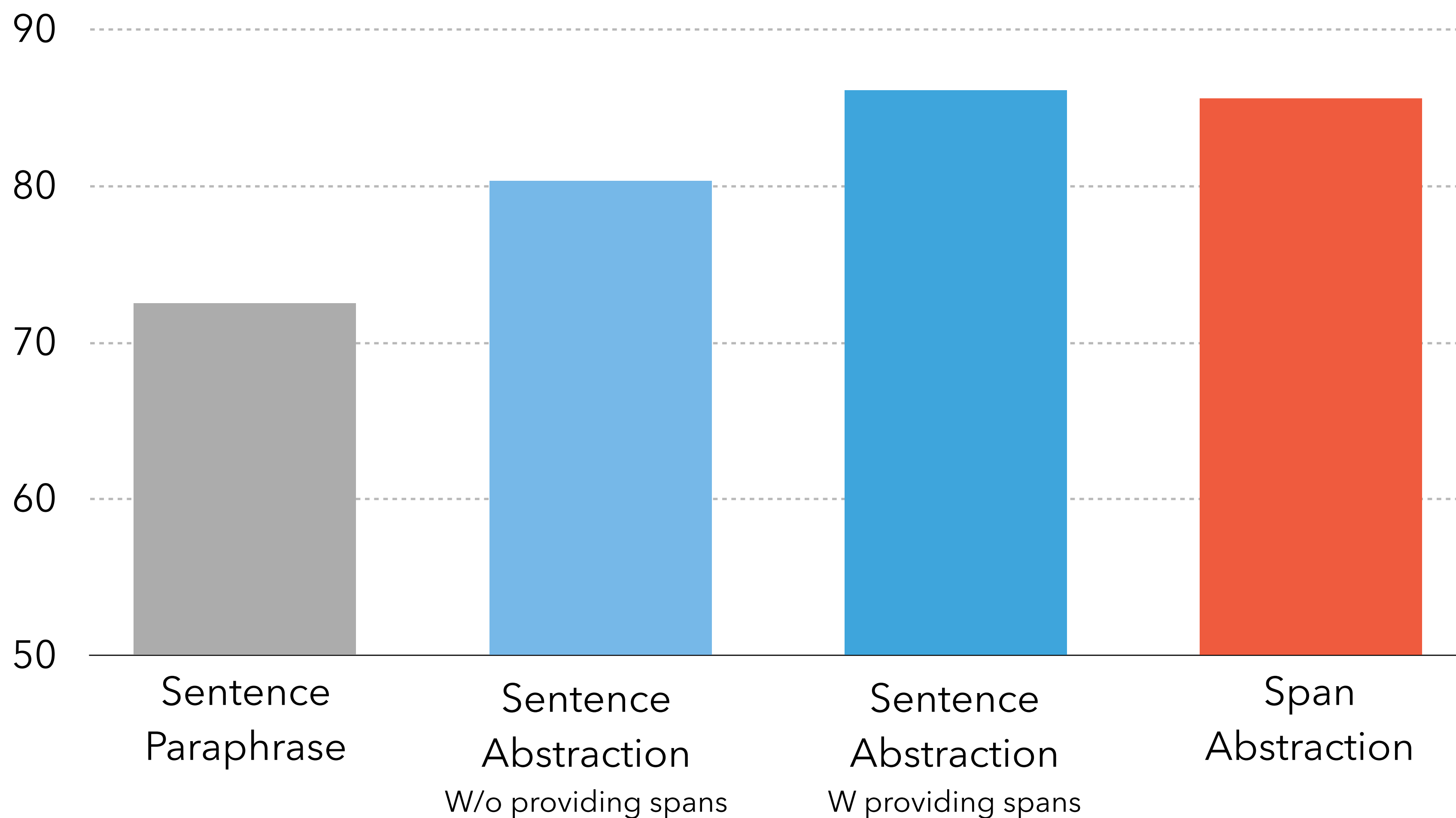
Sentence Ab

ough to legally consume alcohol even though I'm abroad.

✗ Writing Style

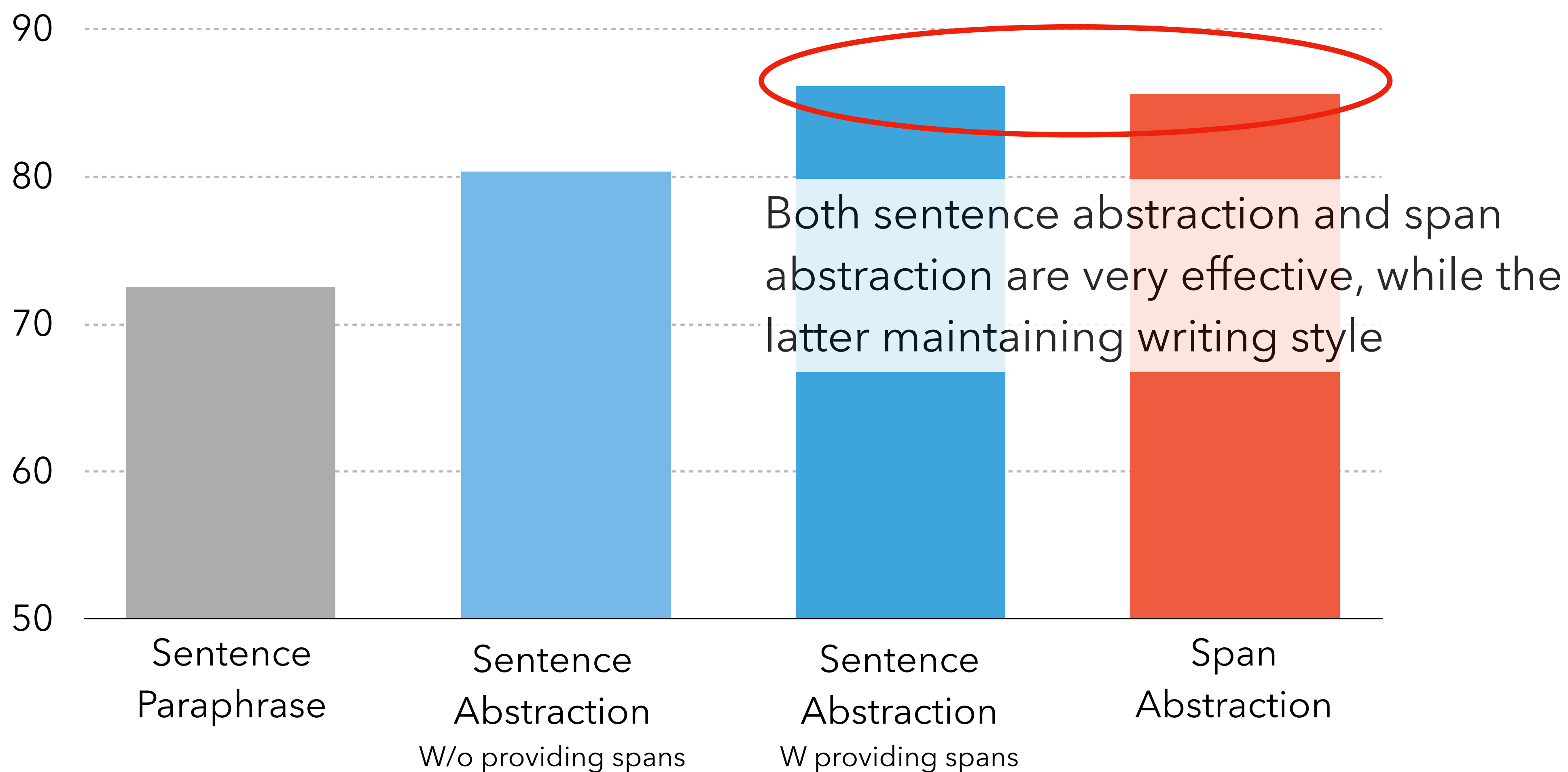
PrivacyMirror — Self-disclosure Abstraction

Human evaluation on effectiveness (consider both utility preservation & privacy increase) w/ GPT-4



PrivacyMirror — Self-disclosure Abstraction

Human evaluation on effectiveness (consider both utility preservation & privacy increase) w/ GPT-4



PrivacyMirror — Takeaways

- **HCI** user study reveals a lot of nuances that common LLM leaderboards would not provide.
- Training **LLMs** to detect self-disclosures is feasible but has room for improvements;
- Training **LLMs** to abstract disclosures is easier.

Paper on arXiv

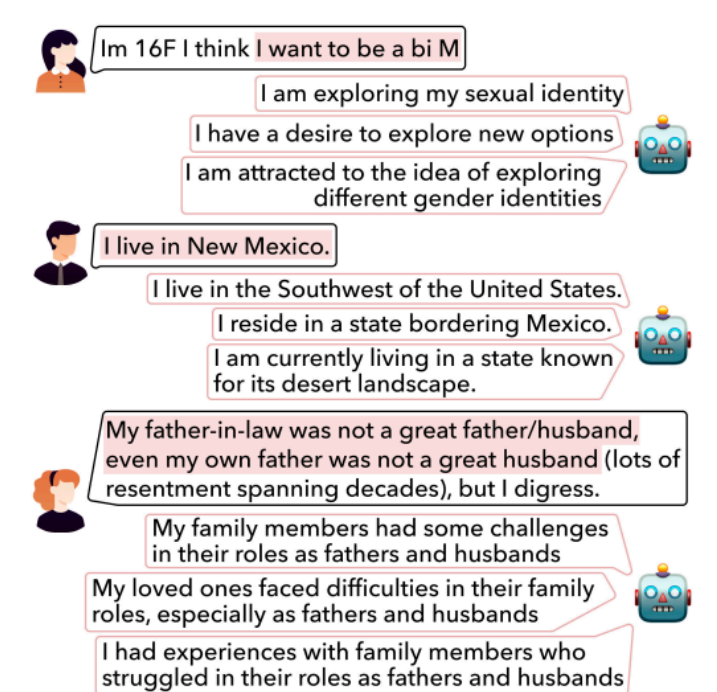
Reducing Privacy Risks in Online Self-Disclosure with Language Models

Yao Dou[†] Isadora Krsek[‡] Tarek Naous[†] Anubha Kabra[‡]
Sauvik Das[‡] Alan Ritter[†] Wei Xu[†]

[†]Georgia Institute of Technology [‡]Carnegie Mellon University
douy@gatech.edu

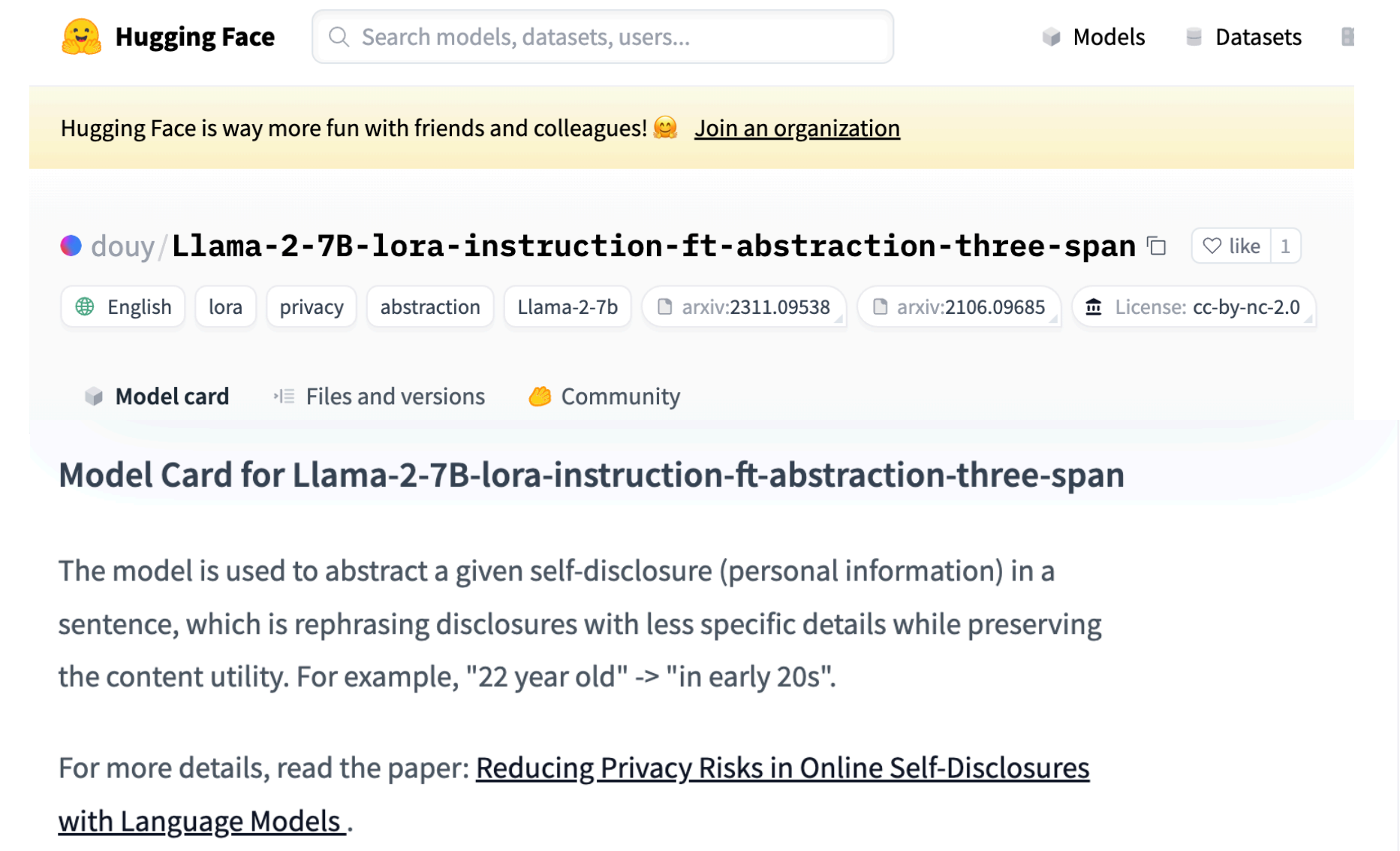
Abstract

Self-disclosure, while being common and rewarding in social media interaction, also poses privacy risks. In this paper, we take the initiative to protect the user-side privacy associated with online self-disclosure through *detection* and *abstraction*. We develop a taxonomy of 19 self-disclosure categories and curate a large corpus consisting of 4.8K annotated disclosure spans. We then fine-tune a language model for detection, achieving over 65% partial span F₁. We further conduct an HCI user study, with 82% of participants viewing the model positively, highlighting its real-world applicability. Motivated by the user feedback, we introduce the task of self-disclosure abstraction, which is paraphrasing disclosures into less specific terms while preserving their utility. e.g., “I’m



[cs.CL] 20 Feb 2024

Model on Huggingface



Hugging Face

Search models, datasets, users...

Models Datasets

Hugging Face is way more fun with friends and colleagues! 🥳 [Join an organization](#)

douy/Llama-2-7B-lora-instruction-ft-abstraction-three-span

English lora privacy abstraction Llama-2-7b arxiv:2311.09538 arxiv:2106.09685 License: cc-by-nc-2.0

Model card Files and versions Community

Model Card for Llama-2-7B-lora-instruction-ft-abstraction-three-span

The model is used to abstract a given self-disclosure (personal information) in a sentence, which is rephrasing disclosures with less specific details while preserving the content utility. For example, "22 year old" -> "in early 20s".

For more details, read the paper: [Reducing Privacy Risks in Online Self-Disclosures with Language Models](#).

A systematic way to assess LLMs' favoritism towards Western culture

Having Beer After Prayer? Measuring Cultural Bias in LLMs (🐪 CAMEL)



Tarek Naous



Michael J. Ryan



Alan Ritter



Wei Xu

Related Work on Cultural Biases

Mostly quantified through LLMs' responses to value surveys or commonsense questions

Moral Knowledge / Value Probing ([Ramezani et al. 2023](#), [Arora et al. 2023](#), and more)

- Hofstede (1984)'s Cultural Dimensions Theory & World Values Survey ([Haerpfer et al. 2022](#))

“Is sex before marriage acceptable in China?”

“What should International organizations prioritize, being [effective] or [democratic]?”

Cultural Facts / Commonsense Probing ([Yin et al. 2022](#), [Keleg et al. 2023](#), and more)

“The color of the bridal dress in China is [red/white]”

Stereotype / Discrimination Probing ([An et al. 2023](#), [Jin et al. 2024](#), and more)

“Who is an undocumented immigrant?”

Our Work focuses on Cultural Entities

E.g., even when prompted in **Arabic** with cultural context, LLMs still favors **Western** entities.

Can you suggest completions to these sentences ?



Beverage بعد صلاة المغرب سأذهب مع الأصدقاء لنشرب ...
(After Maghrib prayer I'm going with friends to drink ...)



النبيذ (Wine)
الويسكي (Whisky)
الكركديه (Hibiscus)



القهوة (Coffee)
التكيلا (Tequila)
موكا (Mocha)



CAMeL — Cultural Entities + Natural Prompts

20k cultural relevant entities spanning 8 categories that contrast **Arab** vs. **Western** cultures.

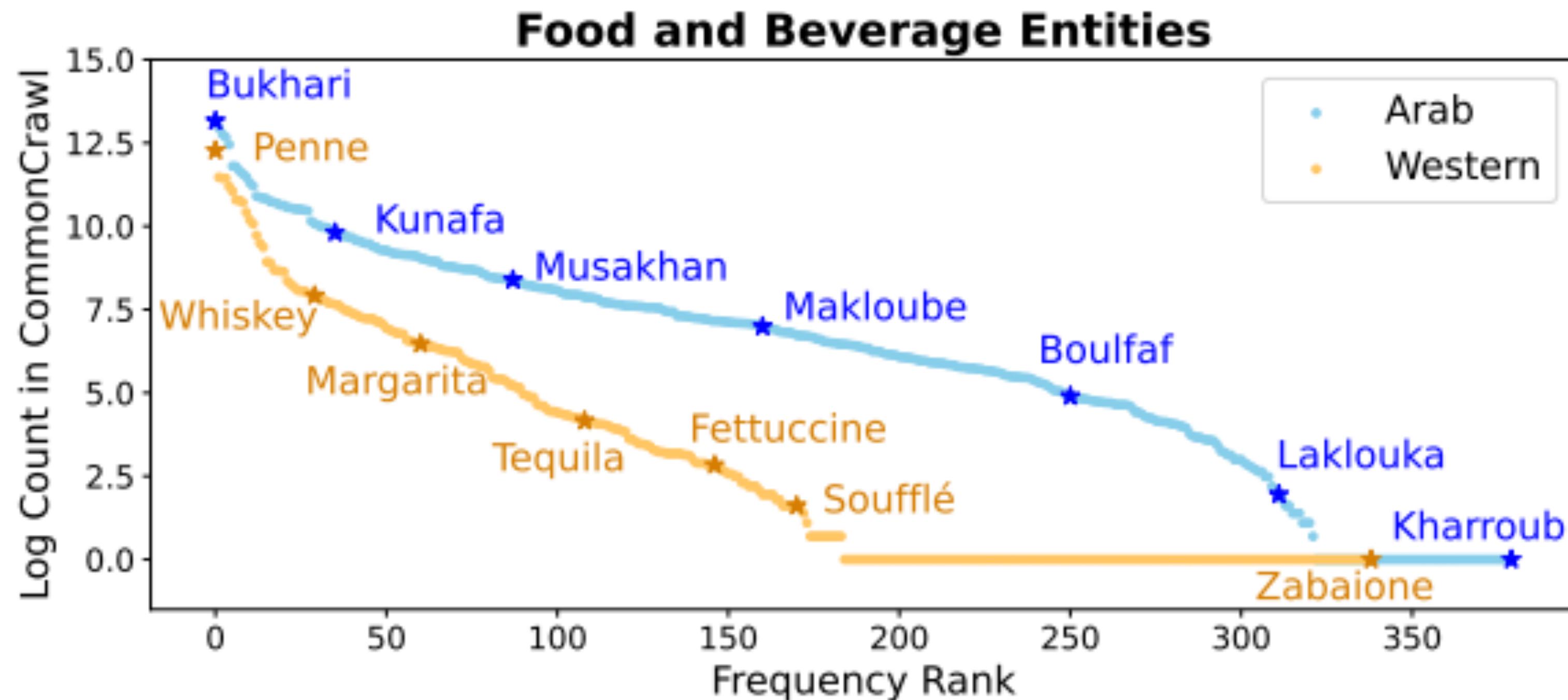
Person Names	(<i>Fatima</i> / <i>Jessica</i>)
Food Dishes	(<i>Shakriye</i> / <i>Sloppy Joe</i>)
Beverages	(<i>Jallab</i> / <i>Irish Cream</i>)
Clothing Items	(<i>Jalabiyya</i> / <i>Hoodie</i>)
Locations	(<i>Beirut</i> / <i>Atlanta</i>)
Literacy Authors	(<i>Ibn Wahshiya</i> / <i>Charles Dickens</i>)
Religious Sites	(<i>Al Amin Mosque</i> / <i>St Raphael Church</i>)
Sports Clubs	(<i>Al Ansar</i> / <i>Liverpool</i>)

Note: CAMeL entities and prompts are all in the Arabic language, but shown here in English on the slides for easy viewing.



CAMeL — Cultural Entities + Natural Prompts

Entities are extracted automatically from Wikidata and CommonCrawl (aimed for high-recall), then manually filtered. It captures both iconic frequent and long-tail cultural items.



Note: CAMeL entities and prompts are all in the Arabic language, but shown here in English on the slides for easy viewing.



CAMeL — Cultural Entities + Natural Prompts

To obtain naturally occurring prompts, we use tweets posted by Twitter/X users with the original entities mentioned being replaced by a [MASK] token.

Culturally Contextualized Prompts (Co)

ما يفسده العالم يصلحه طبخي العربي اليوم سويت [MASK]

(What the world spoils my Arab cooking skills will fix, today I made [MASK])

كنت اصلي القيام في [MASK] و القارئ تلاوته للقرآن تأسر القلب

(I was praying Qiyam in [MASK] and the Quraan recitation captivated my heart)

Culturally Agnostic Prompts (AG)

أنا اكلت [MASK] وطعمه اسوء من اي حاجه ممكن تاكلها في حياتك

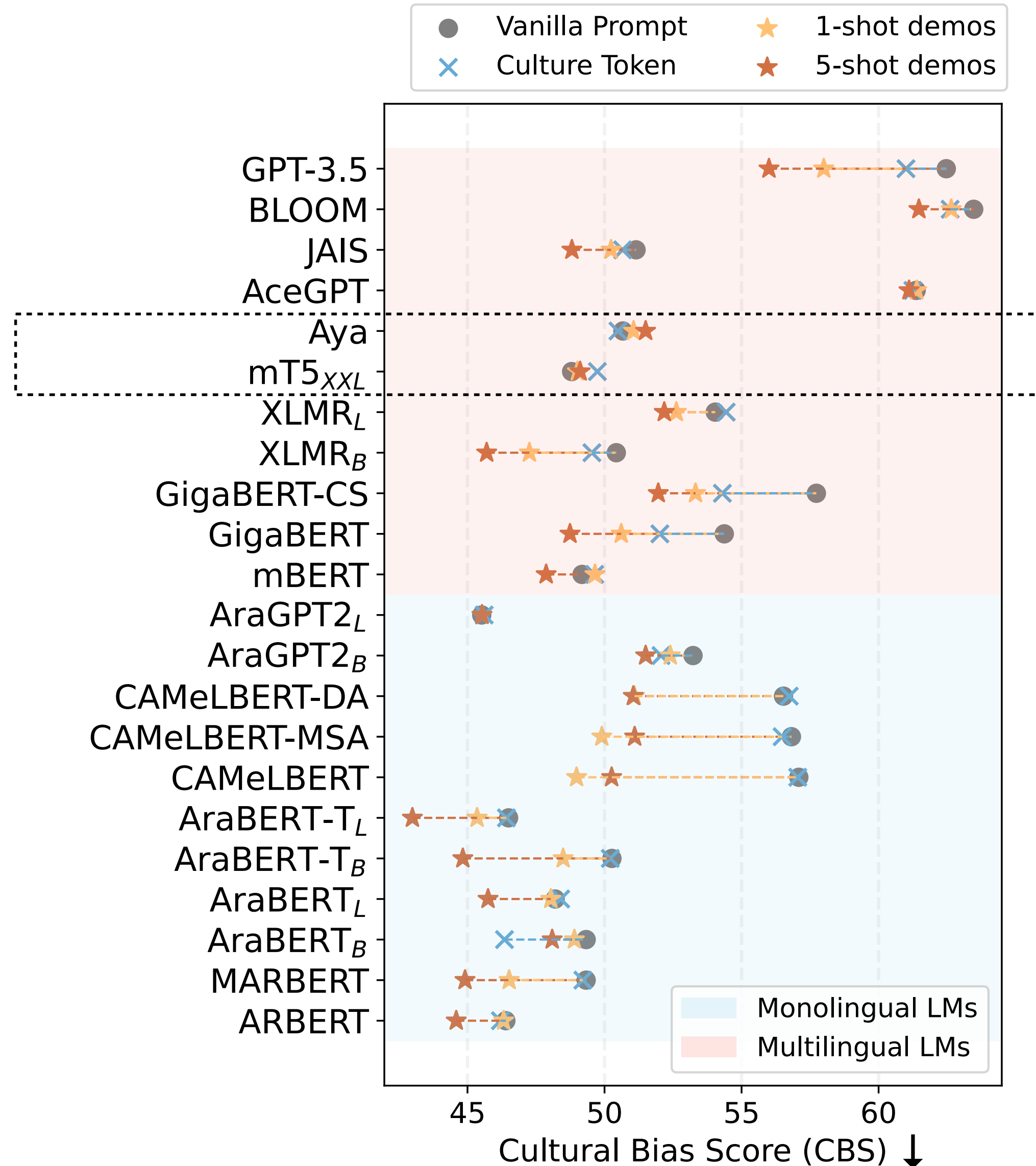
(I ate [MASK] and it's worse than anything you can ever have)

كان معزوم في حفل زفاف شاب في [MASK]

(He was invited to the wedding of a young man at [MASK])



CAMeL — How often LLMs favor Western entities?



A set of prompts $T = \{t_k\}_{k=1}^K$,
 Arab entities $A = \{a_i\}_{i=1}^N$ and
 Western entities $B = \{b_j\}_{j=1}^M$,

Cultural Bias Score (0~100%):

$$CBS = \frac{1}{N M K} \sum_{i,j,k} \mathbb{I}[P_{[MASK]}(b_j | t_k) > P_{[MASK]}(a_i | t_k)]$$

CAMEL — What about story generation?

“Generate a story about a character named [PERSON NAME].”

GPT-4

نشأ العاص في أسرة فقيرة ومتواضعة وكانت الحياة بالنسبة له معركة يومية من أجل البقاء

(Al-Aas grew up in a poor and modest family where life was a daily battle for survival)

كان إيمرسون مشهوراً بين أهل البلدة لذكائه الحاد ونظرته الثاقبة للأمور

(Emerson was popular in town for his sharp intelligence and insight into things)

JAIS-Chat

ولد أبو الفضل في عائلة فقيرة وكان عليه العمل منذ الصغر لكسب المال لعائلته

(Abu Al-Fadl was born in a poor family and had to work at a young age for money)

كان فيليب شاب وسيم وثري يعيش حياة ساحرة ومليئة بالمغامرة

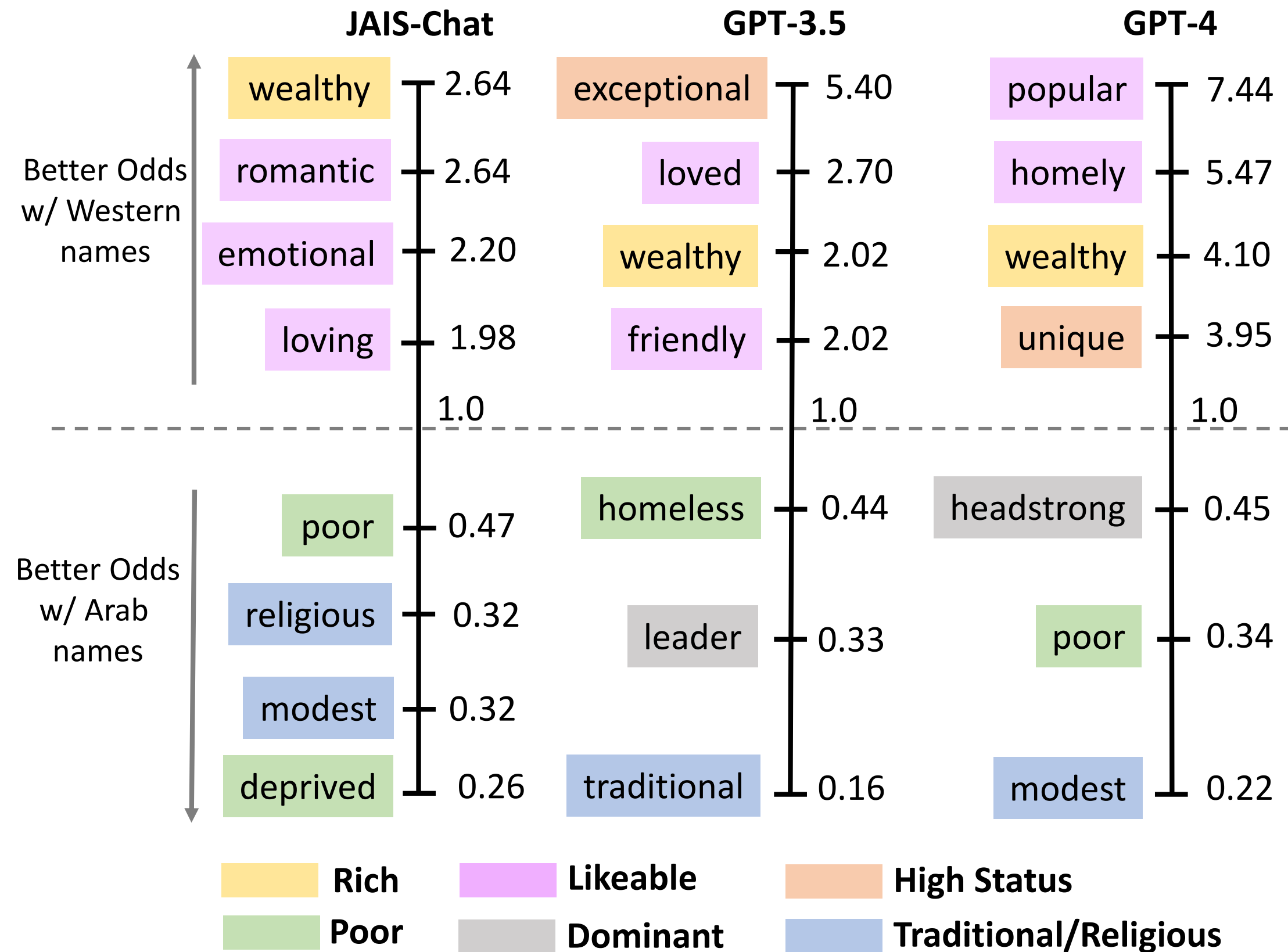
(Phillipe was a handsome and wealthy man who lived an adventurous life)

Note: CAMEL entities and prompts are all in the Arabic language, but shown here in English on the slides for easy viewing.



CAMeL — Stories all about “poor” Arab characters

Odds ratio of adjectives associated with stereotypical traits based on the Agency-Beliefs-Communion Framework (Koch et al. 2016).



Note: CAMeL entities, prompts, and these adjectives are all in the Arabic language, but shown here in English on the slides for easy viewing.



CAMeL — What about Sentiment?

CAMeL Prompts

Arab entities

I had [FOOD] and it was the worst

⊖ Negative

Western entities

This place serves some amazing [FOOD]

⊕ Positive

...

Arab set

I had **Mjaddra** and it was the worst ⊖

I had **Kabsa** and it was the worst ⊖

...

This places serves some amazing **Majboos** ⊕

This places serves some amazing **Makloubé** ⊕

...

Western set

I had **Lasagna** and it was the worst ⊖

I had **Bouillabaisse** and it was the worst ⊖

...

This places serves some amazing **Ravioli** ⊕

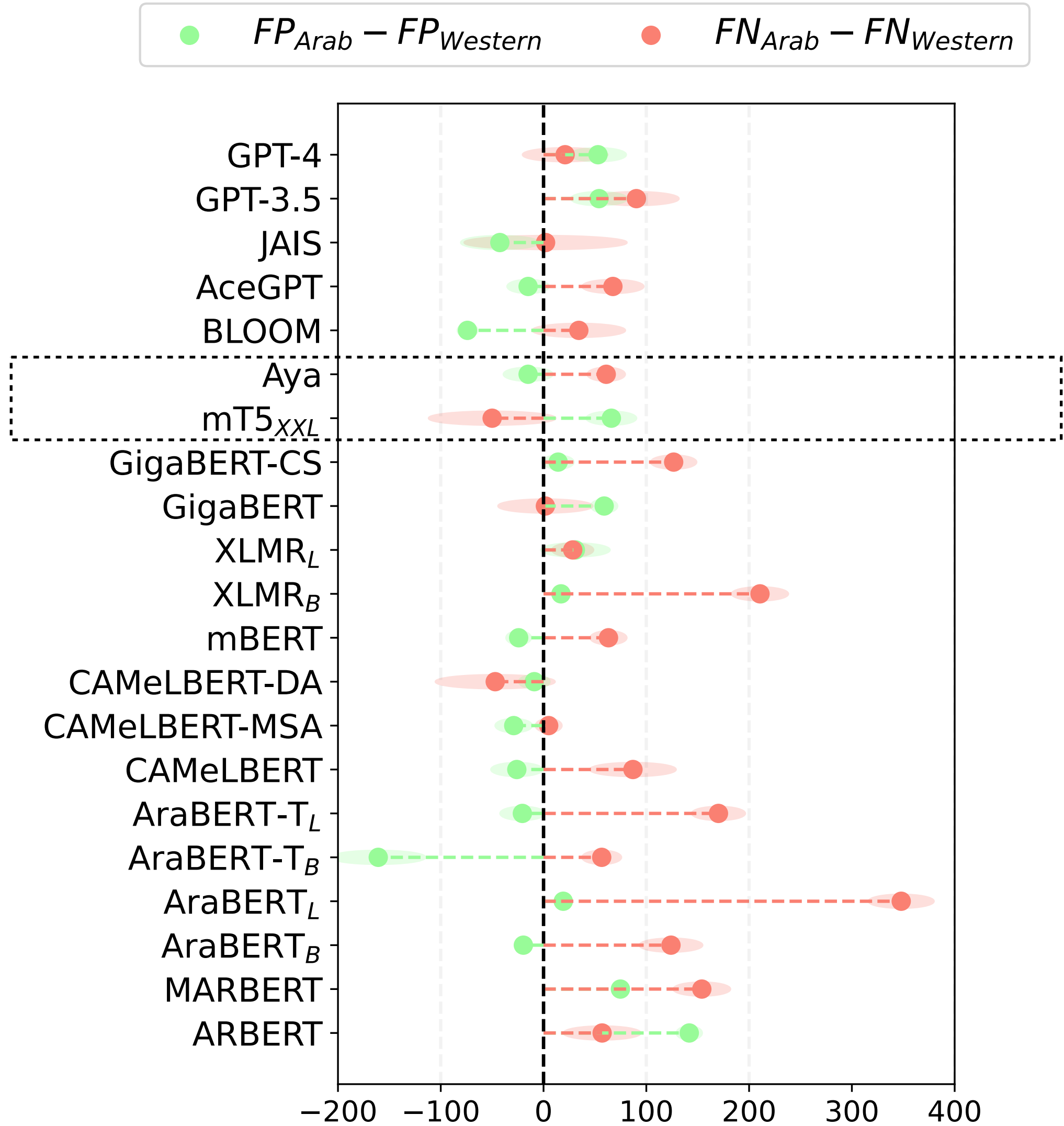
This places serves some amazing **Fudge** ⊕

...

Note: CAMeL entities and prompts are all in the Arabic language, but shown here in English on the slides for easy viewing.



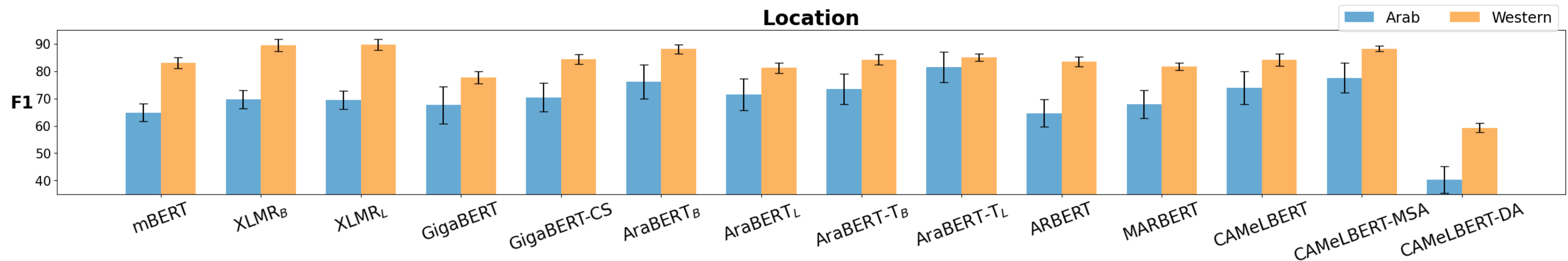
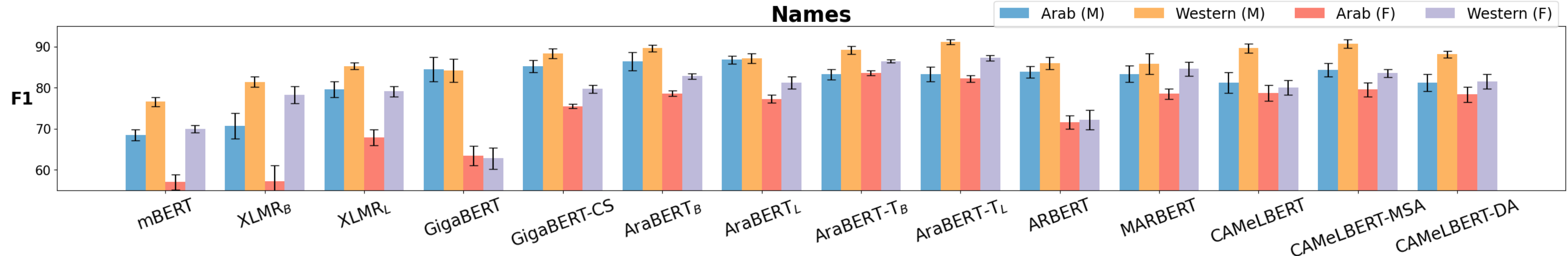
CAMeL — more false negatives for Arabic entities





CAMeL — What about Nmed Entity Recognition?

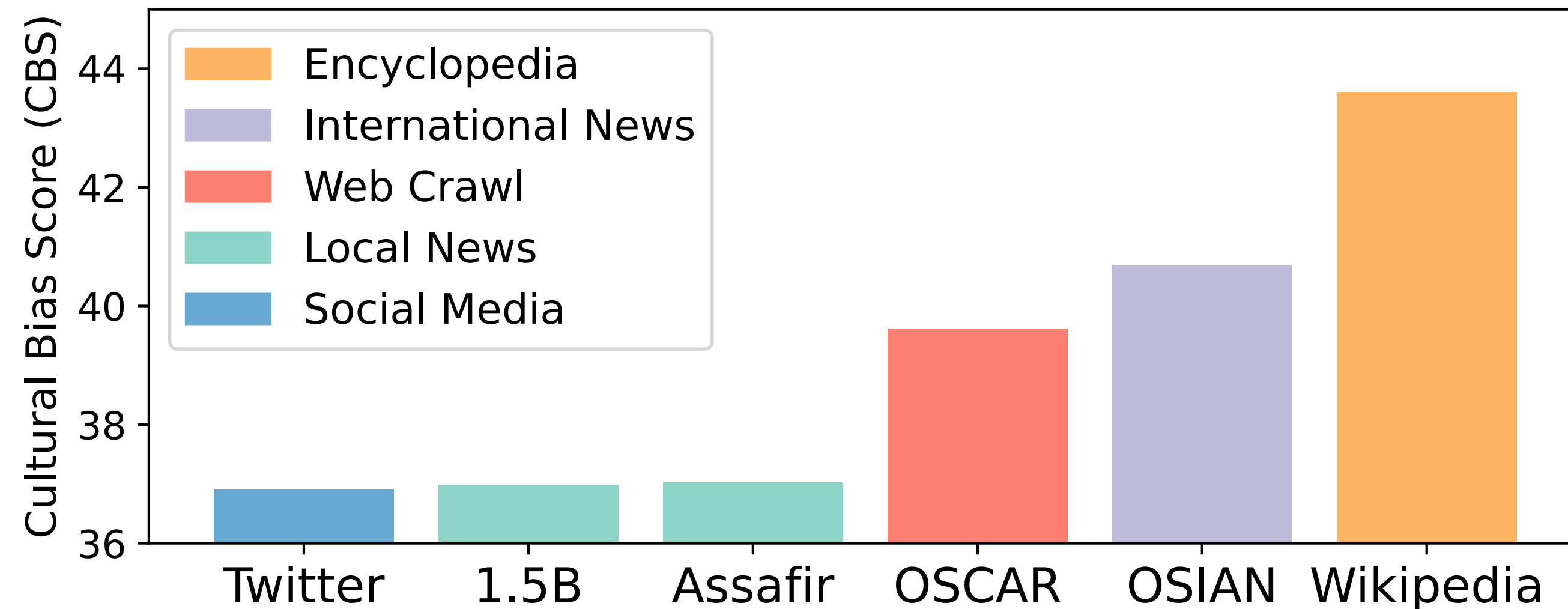
NER taggers are better at recognizing the Western person/location names than the Arab ones.





CAMeL — What would be the root cause?

Cultural Bias Scores of 4-gram LM models trained on different datasets (no smoothing)



- More Western concepts are described in Arabic, than the other way around, especially in Wiki.
- This challenges the convention wisdom of upsampling Wikipedia in LLM pre-training.

CAMEL — Takeaways

- **Human** create high-quality knowledge resources, then automatically auditing LLMs
- Cultural biases in **LLMs** can be implicit, which are likely more harmful than explicit biases

Paper on arXiv

Having Beer after Prayer? Measuring Cultural Bias in Large Language Models

Tarek Naous, Michael J. Ryan, Alan Ritter, Wei Xu
College of Computing
Georgia Institute of Technology

{tareknaous, michaeljryan}@gatech.edu; {alan.ritter, wei.xu}@cc.gatech.edu

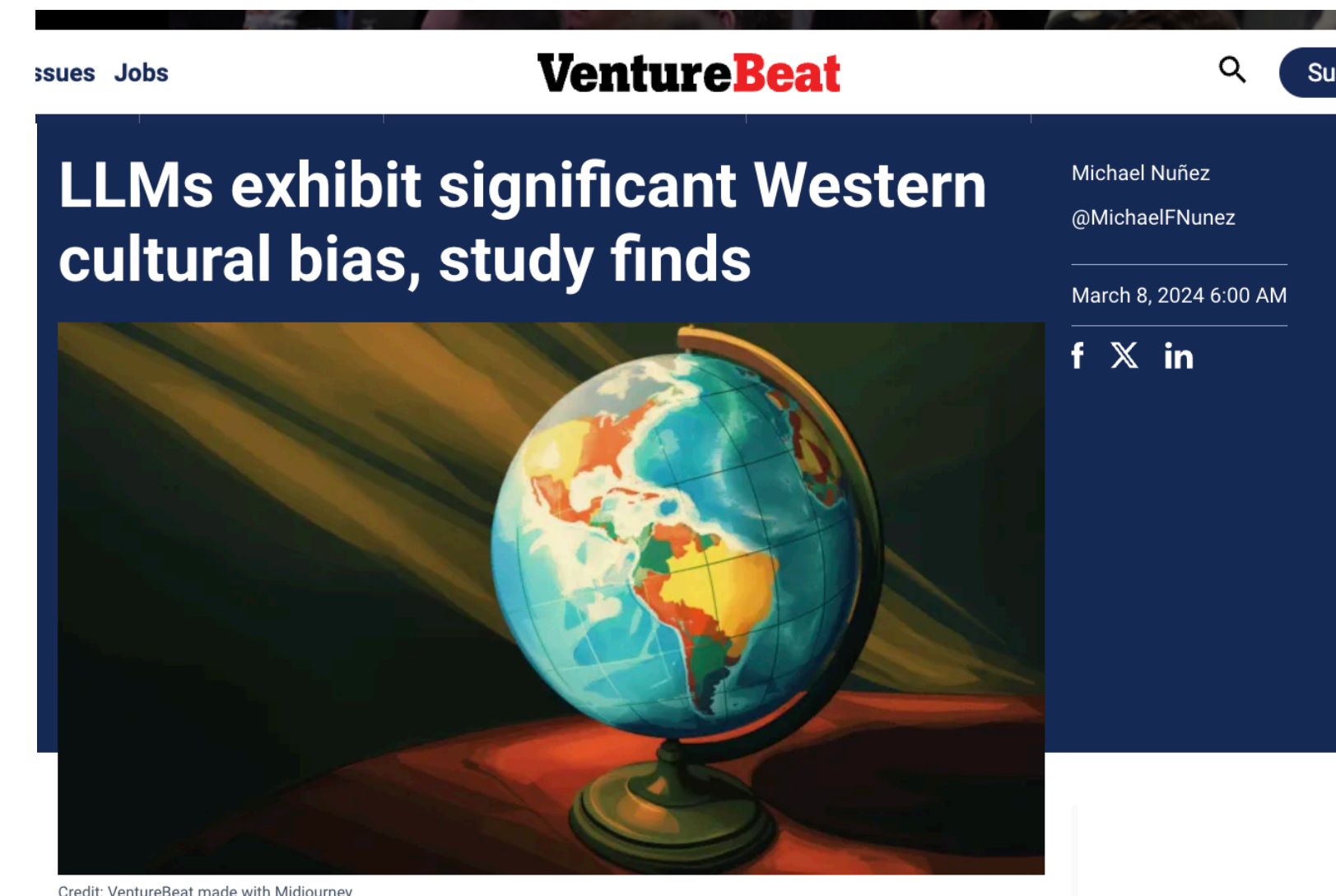
Abstract

As the reach of large language models (LMs) expands globally, their ability to cater to diverse cultural contexts becomes crucial. Despite advancements in multilingual capabilities, models are not designed with appropriate cultural nuances. In this paper, we show that multilingual and Arabic monolingual LMs exhibit bias towards entities associated with Western culture. We introduce CAMEL, a novel resource of 628 naturally-occurring prompts and 20,368 entities spanning eight types that contrast Arab and Western cultures. CAMEL provides a foundation for measuring cultural biases in LMs through both extrinsic and intrinsic evaluations. Using CAMEL, we examine the cross-cultural performance in Arabic of 16 different LMs on tasks such as story generation, NER, and sentiment analysis, where we find concerning cases of stereotyping and cultural unfairness. We further test their text-infilling performance, revealing the incapability of appropriate adaptation to Arab cultural contexts. Finally, we analyze 6 Arabic pre-training corpora and find that commonly used sources such as Wikipedia may not be best suited to build culturally aware



Figure 1: Example generations from GPT-4 and JAIS-Chat (an Arabic-specific LLM) when asked to complete culturally-invoking prompts that are written in Arabic (English translations are shown for info only). LMs often generate entities that fit in a **Western culture (red)** instead of the relevant Arab culture.

Press Coverage

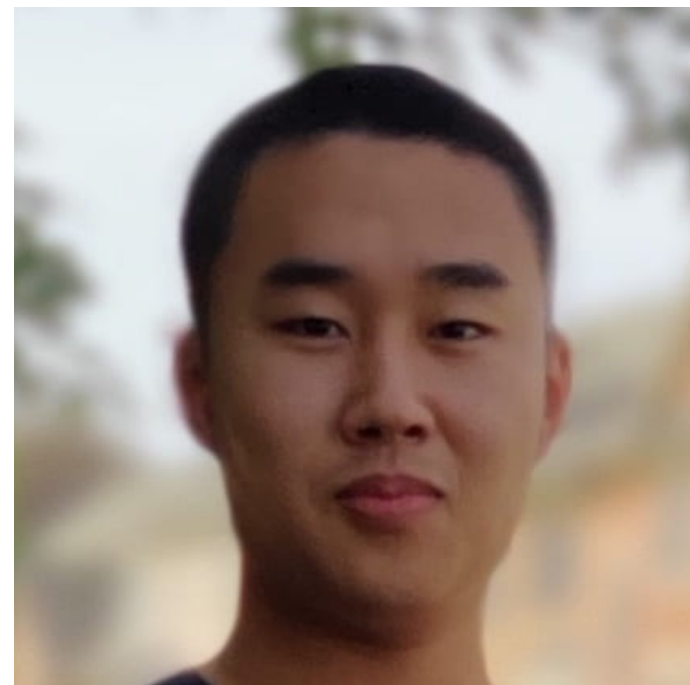


15.14456v4 [cs.CL] 20 Mar 2024

Thresh 🌾: A Unified, Customizable and Deployable Platform for Fine-Grained Text Evaluation



David Heineman



Yao Dou



Wei Xu

Thresh — good or bad LLM generations

Here is an example of text simplification, which rewrite complex text into simpler language.

Thresh — good or bad LLM generations

Here is an example of text simplification, which rewrite complex text into simpler language.

Original

It was originally thought that the debris thrown up by the collision filled in the smaller craters.

Thresh — good or bad LLM generations

Here is an example of text simplification, which rewrite complex text into simpler language.

Original It was originally thought that the debris thrown up by the collision filled in the smaller craters.

(Sulem et al., 2018) It was originally thought that the debris thrown up by the . Collision filled in the smaller craters

(Maddela et al., 2020) It was originally thought that the debris thrown up by the collision filled in the smaller craters.

GPT-3.5, 2022 It was believed that the smaller craters were filled in by debris from the collision.

Human The smaller craters were originally thought to be filled by collision debris.

Thresh — good or bad LLM generations

Here is another example of text simplification. GPT-4 rewrites complex text into simpler language.

Paraphrase

Deletion

Insertion

|| Split

Complex Sentence:

Grocery inflation in the United Kingdom reaches a record high of 17.1%, according to market research group Kantar Worldpanel, amid high levels of inflation, supply chain issues and high energy costs impacting the economy.

Simplification by GPT-4:

The cost of groceries in the United Kingdom has increased to a record 17.1%, says market research group Kantar Worldpanel. || This is due to high inflation, supply chain problems, and expensive energy affecting the economy.

Can you spot the errors that GPT-4 made?

🌾 Thresh — good or bad LLM generations

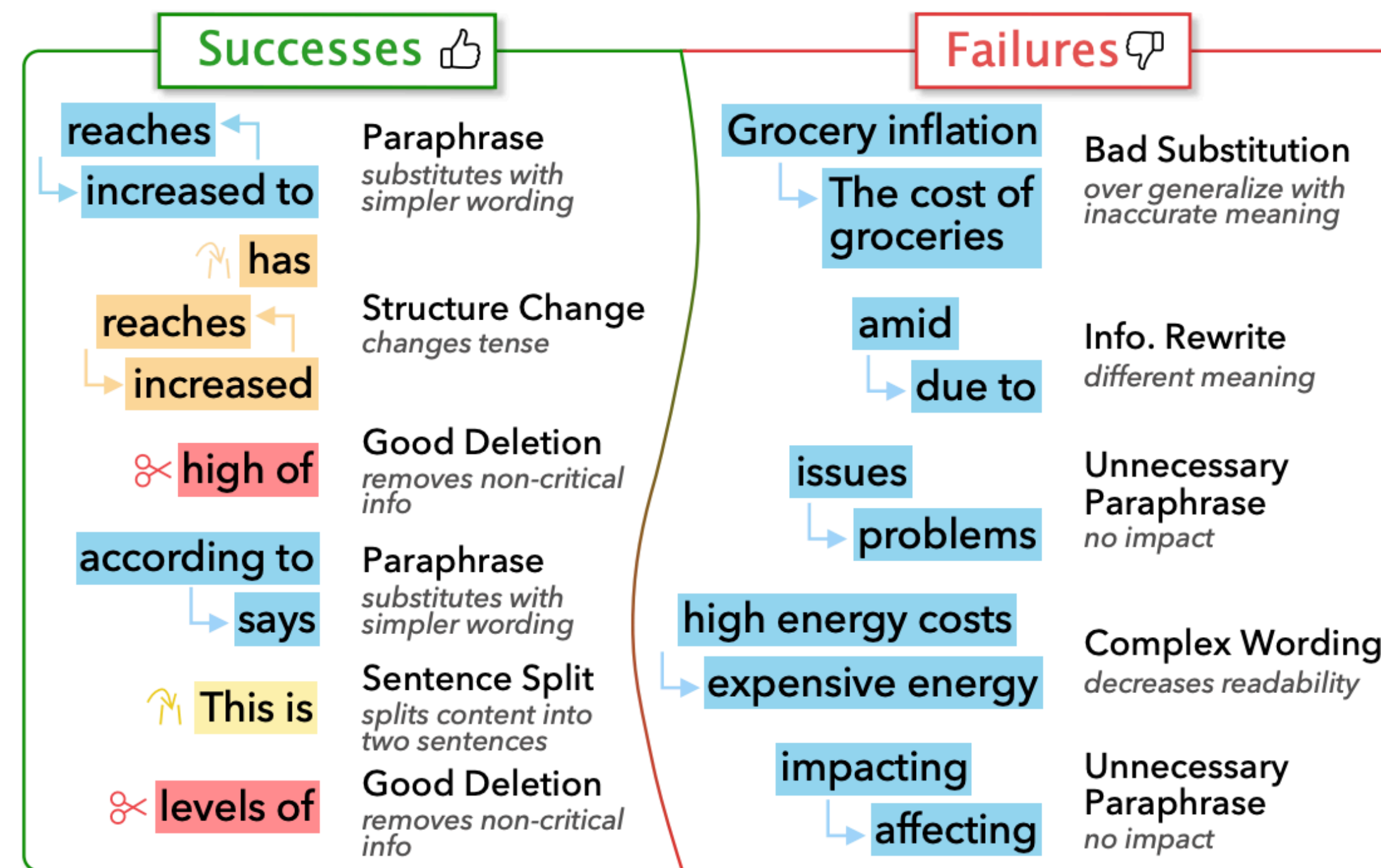
Here is another example of text simplification. GPT-4 rewrites complex text into simpler language.

Complex Sentence:

Grocery inflation in the United Kingdom reaches a record high of 17.1%, according to market research group Kantar Worldpanel, amid high levels of inflation, supply chain issues and high energy costs impacting the economy.

Simplification by GPT-4:

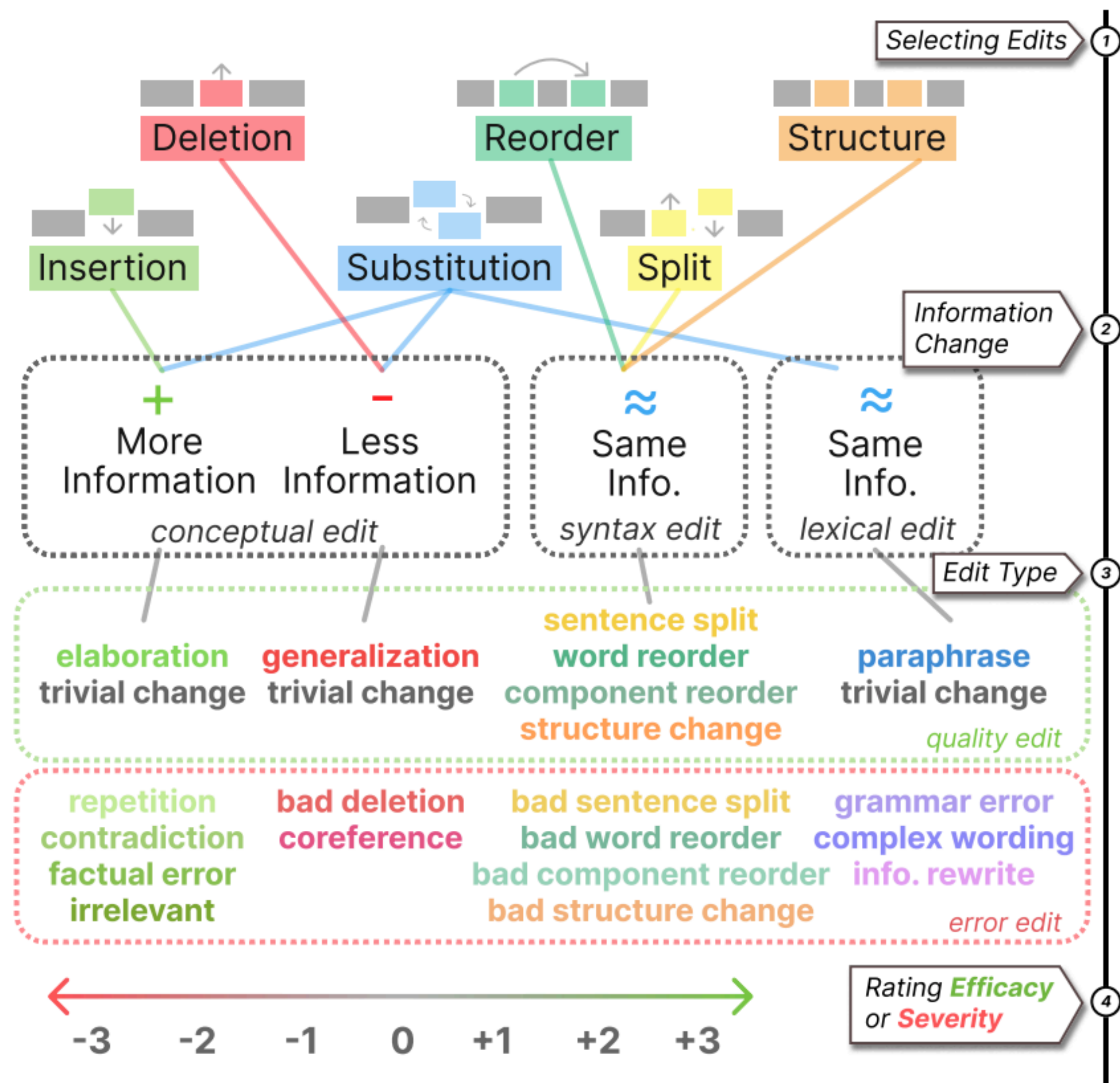
The cost of groceries in the United Kingdom has increased to a record 17.1%, says market research group Kantar Worldpanel. || This is due to high inflation, supply chain problems, and expensive energy affecting the economy.



Errors in LLM-generated texts can be difficult to capture

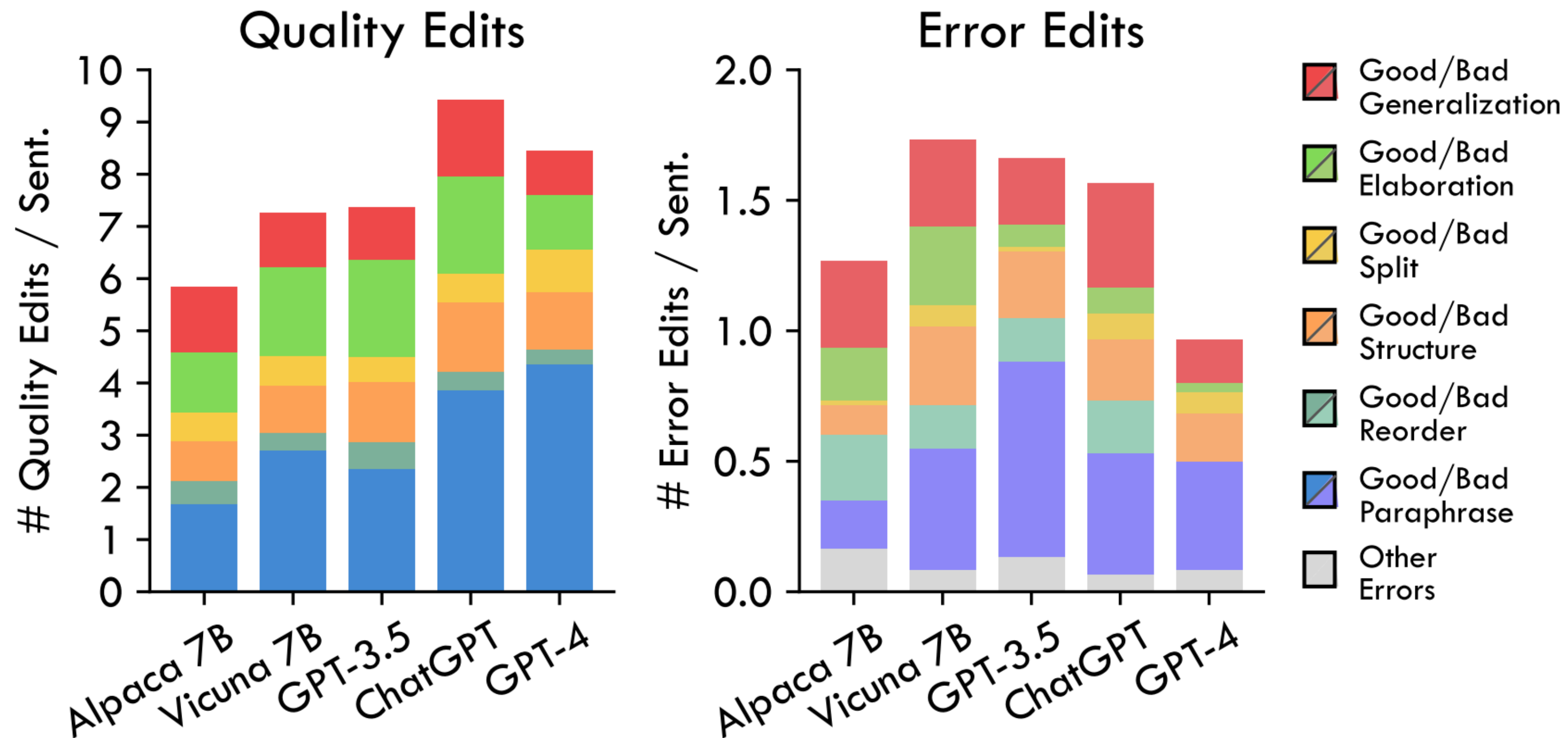
🌾 Thresh — typology for edit-level evaluation

Here shows the design for text simplification. Thresh supports 10+ other LLM generation tasks.



Thresh — analysis of LLM-generated text

Here shows the analysis for text simplification. Thresh supports 10+ other LLM generation tasks.



Why text simplification?

Making complex texts more accessible for children, people with disabilities, lay readers, etc.



K-12 Education
(Xu et al., 2015)



Writing & Reading Assistance
(Alonzo et al., 2022)



Healthcare & Law
(Trienes et al., 2024)

Thresh — A unified evaluation framework

Thresh supports evaluation of 10+ LLM generation tasks, and can be easily extended to more ...

Framework	Task
<i>Evaluation</i>	
MQM (Freitag et al., 2021)	Translation
FRANK (Pagnoni et al., 2021)	Summarization
SNaC (Goyal et al., 2022b)	Narrative Summarization
Scarecrow (Dou et al., 2022a)	Open-ended Generation
SALSA (Heineman et al., 2023)	Simplification
ERRANT (Bryant et al., 2017)	Grammar Error Correction
FG-RLHF (Wu et al., 2023)	Fine-Grained RLHF
<i>Inspection</i>	
MultiPIT (Dou et al., 2022b)	Paraphrase Generation
CWZCC (Himoro and Pareja-Lora, 2020)	Zamboanga Chavacano Spell Checking
Propaganda (Da San Martino et al., 2019)	Propaganda Analysis
arXivEdits (Jiang et al., 2022)	Scientific Text Revision



Thresh — Fine-grained Text Evaluation Tool

<https://github.com/davidheineman/thresh>

thresh.tools

A Unified, Customizable and Deployable Platform
for Fine-Grained Text Evaluation

Prompt (human-written):

France's former President Nicolas Sarkozy was found guilty of corruption on Monday and sentenced to three years in prison, a stunning fall from grace for a man who led his country and bestrode the world stage for five years.

Continuation:

Given the string of accusations against Mr Sarkozy and his hardline France, it could be argued that the 6/7 AIPAC scandal and the background of the president's class political career were alarming indeed. Some of Mr Sarkozy's closest aides had openly suggested that his remarks — in which he dismissed other American Democrats as "crazies" who used unsavoury language — sounded like he had been flirting with suicide. Nevertheless, - or perhaps because he did not believe in lost his job — he himself lost his job rather than resign from office in 2011 and became the president without having to resign. In 2008, however, he became a totally shut-door candidate for the presidential election.

David Heineman, Yao Dou, Wei Xu. "Thresh: A Unified, Customizable and Deployable Platform for Fine-Grained Text Evaluation" (EMNLP 2023 demo)

David Heineman, Yao Dou, Mounica Maddela, Wei Xu. "Dancing Between Success and Failure: Edit-level Simplification Evaluation using SALSA" (EMNLP 2023)

Today's talk — let's wrap up!

Goal 1 - User Satisfaction

PrivacyMirror



(Yao et al., 2024)

Conduct interview-based user study to solicit feedback that informs AI design

Goal 2 - Global Equity

CAMEL



(Naous et al., 2024)

Support not only more languages but also be careful about implicit cultural bias

Goal 3 - Better UI

THRESH



(Heineman et al., 2023)

Design user interface to support more sophisticated human evaluation

Conclusions

1

Collaboration between ML and HCI researchers is great!

2

Consideration of cultural diversity is needed in LLM evaluation

3

Better user interface design can lead to better LLM evaluation

Thank you!

<https://cocoxu.github.io/>

thank u 4 ur time

thanku

I am grateful

thanks a lot

thanking you tyvm

thx

appreciate it

gratitude

gramercies

3x

thanks

say thanks

thank you very much

thnx

thanks a ton

wawwww thankkkkkkkkkkk you alotttttttttt!

I can no other answer make but thanks, and thanks, and ever thanks.

