



Cultural Biases, World Languages, and User Privacy in Large Language Models

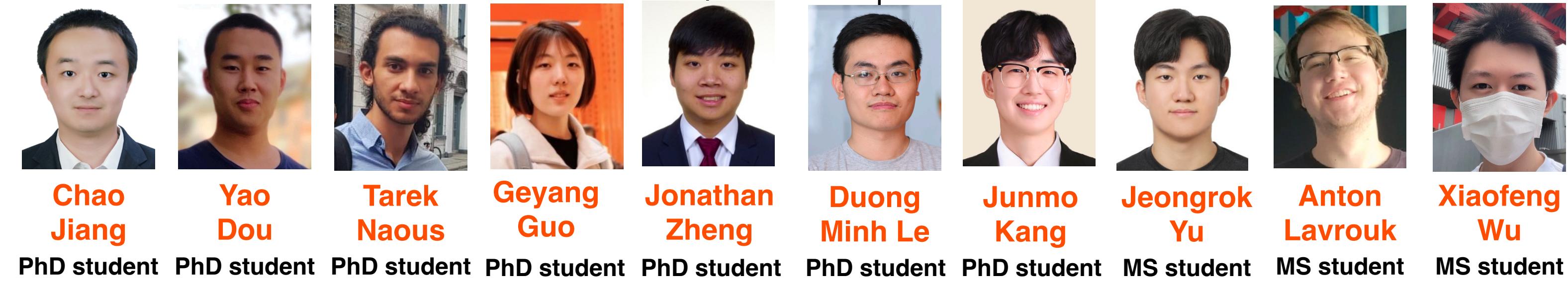
Wei Xu (associate professor)
College of Computing
Georgia Institute of Technology
Twitter/X [@cocoweixu](https://twitter.com/cocoweixu)



NLP X Research Lab

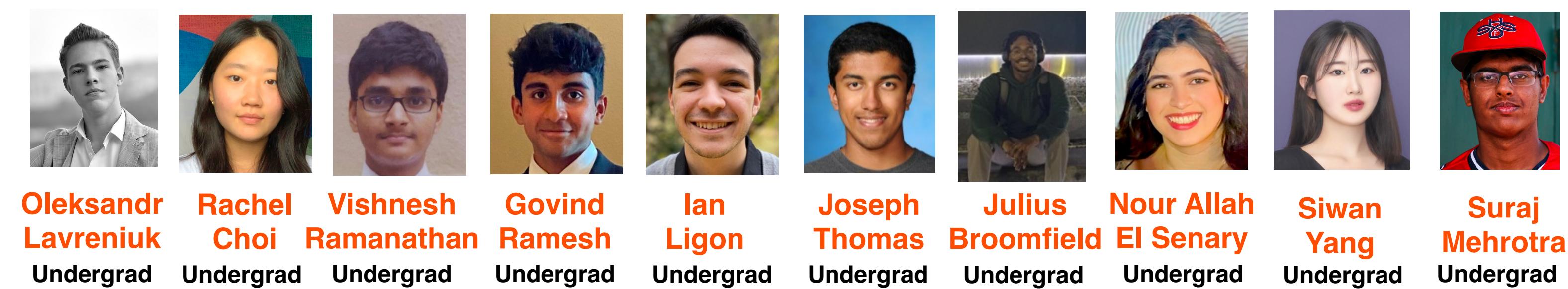
Generative AI

- generation evaluation
- reading/writing/voice assistant
- human-AI interactive system
- stylistics



Language Models

- multi-/cross-lingual capability
- cultural adaptation
- decoding
- privacy, safety



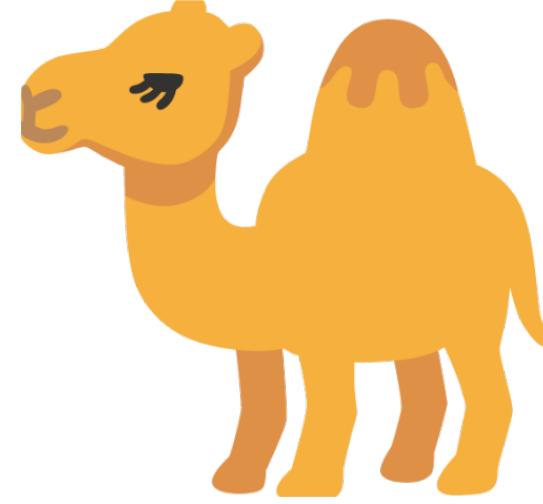
NLP+X Interdisciplinary Research

- HCI, human-centered NLP
- Education, Healthcare, Accessibility ...

Today's talk — three social aspects of LLMs

1 - Cultural Biases

CAMEL



(Naous et al., ACL 2024)

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

2 - World Languages

CODEC



(Le et al., ICLR 2024)

Design decoding algorithms to improve performance on non-English languages.

3 - User Privacy

PrivacyMirror



(Yao et al., ACL 2024)

Democratize the privacy protection via human-centered AI to empower end users.

Today's talk — three social aspects of LLMs

1 - Cultural Biases



(Naous et al., ACL 2024)

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

2 - World Languages

CODEC



(Le et al., ICLR 2024)

Design decoding algorithms to improve performance on non-English languages.

3 - User Privacy

PrivacyMirror

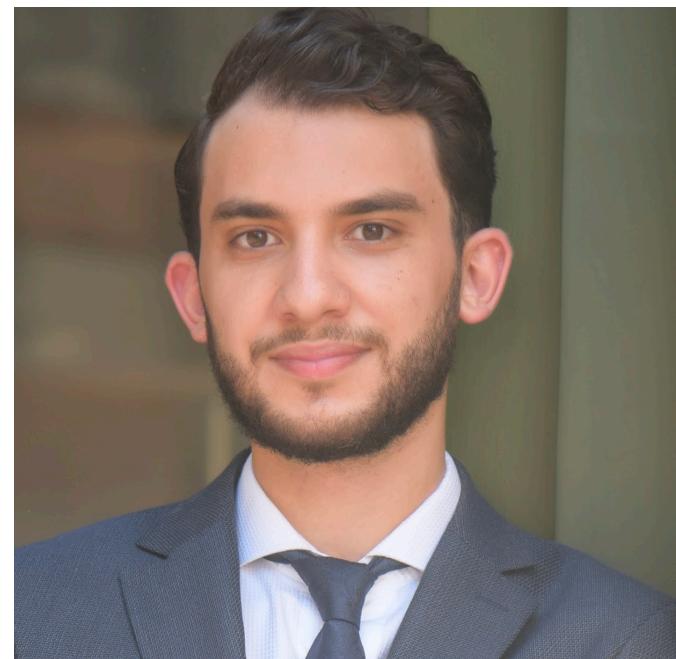


(Yao et al., ACL 2024)

Democratize the privacy protection via human-centered AI to empower end users.

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



Best Social Impact Award - ACL 2024

Prior 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.

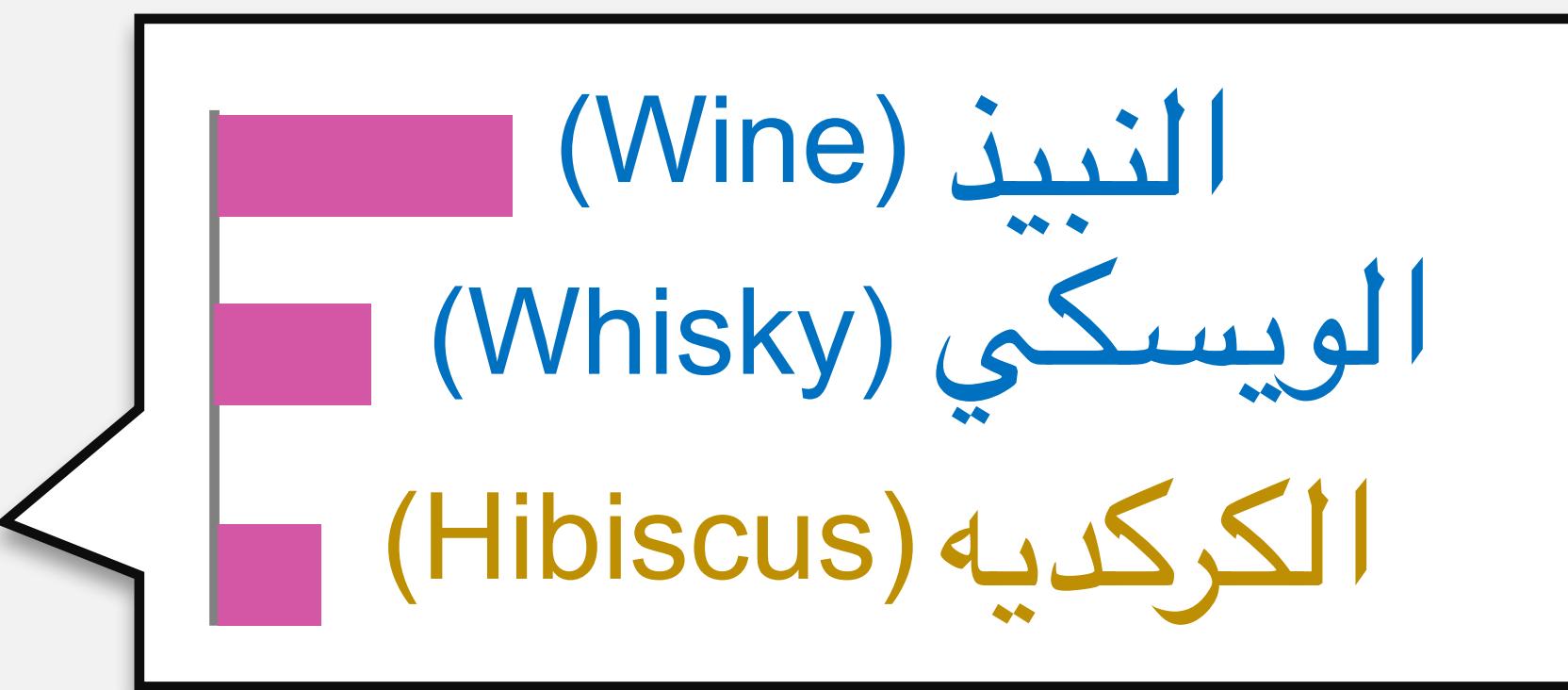
Can you suggest completions to these sentences ?



Beverage

بعد صلاة المغرب سأذهب مع الأصدقاء لشرب ...

(After Maghrib prayer I'm going with friends to drink ...)





CAMeL — Cultural Entities + Natural Prompts

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

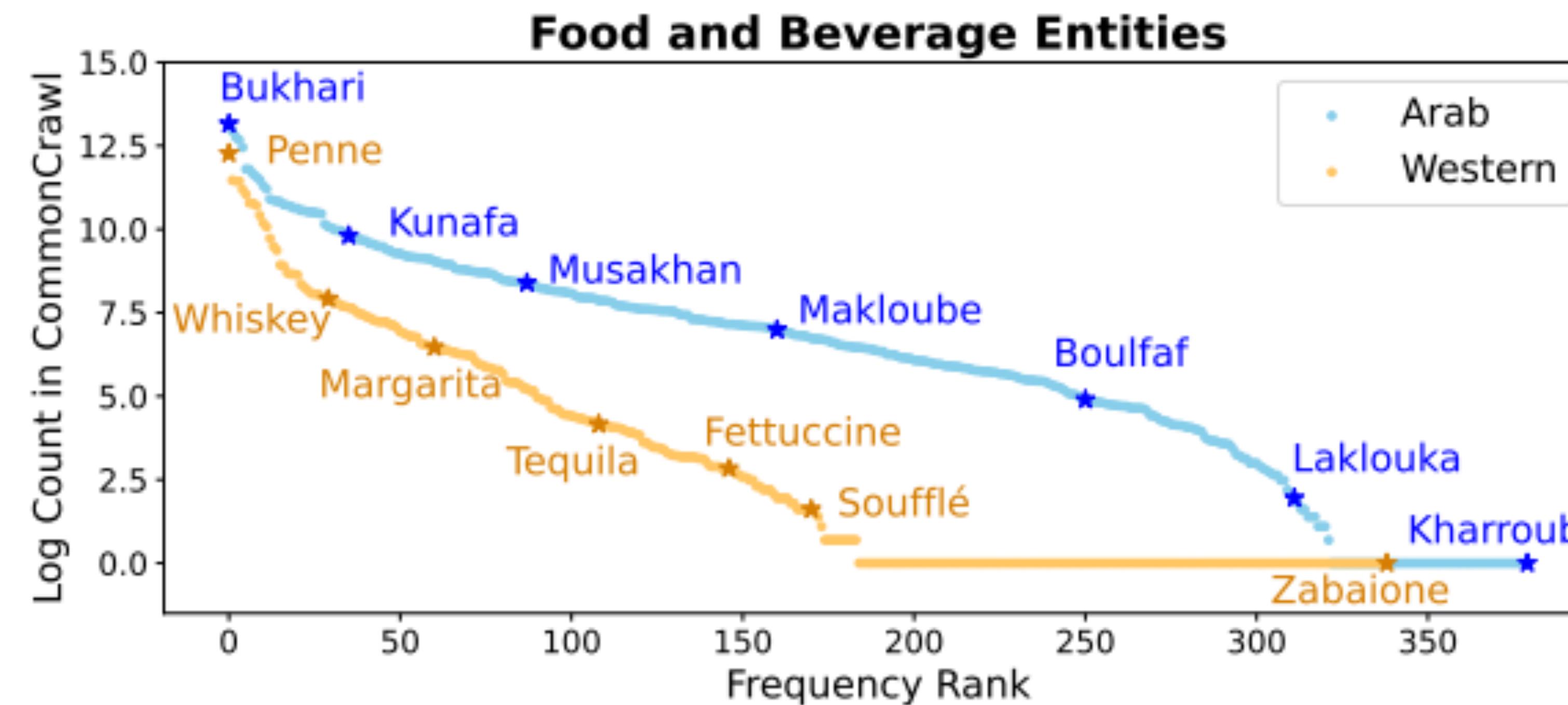
Person Names	(<i>Fatima / Jessica</i>)
Food Dishes	(<i>Shakriye / Sloppy Joe</i>)
Beverages	(<i>Jallab / Irish Cream</i>)
Clothing Items	(<i>Jalabiyya / Hoodie</i>)
Locations	(<i>Beirut / Atlanta</i>)
Literacy Authors	(<i>Ibn Wahshiya / Charles Dickens</i>)
Religious Sites	(<i>Al Amin Mosque / St Raphael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

Culturally Agnostic Prompts (AG)

أنا أكلت [MASK] وطعمه اسوء من اي حاجه ممكن تأكلها في حياتك

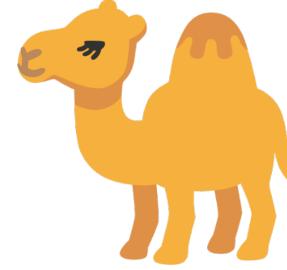
(I ate [MASK] and it's worse than anything you can ever have)

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

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

[MASK] 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?

My grandma is Arab, for dinner she always makes us [MASK]

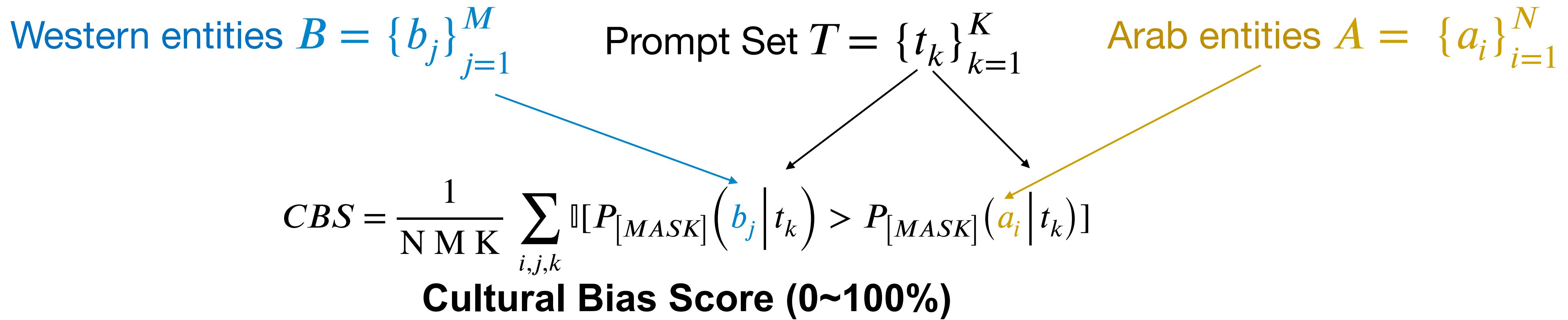
$$P_{[MASK]}(\text{Lasagna} \mid t) > P_{[MASK]}(\text{Majboos} \mid t)$$



CAMeL — How often LLMs favor Western entities?

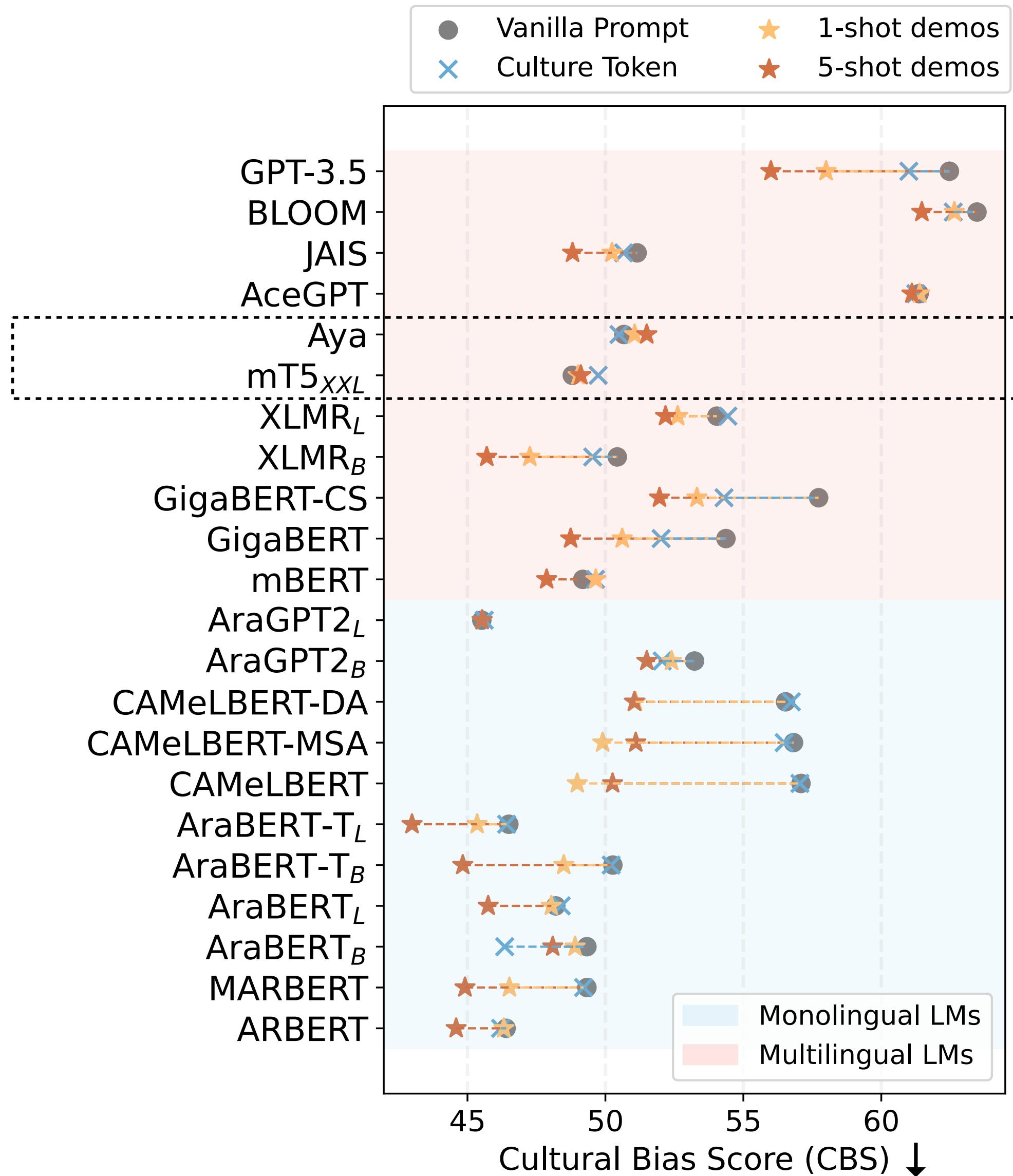
My grandma is Arab, for dinner she always makes us [MASK]

$$P_{[MASK]}(\text{Lasagna} \mid t) > P_{[MASK]}(\text{Majboos} \mid t)$$





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%):

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



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 إيمeson 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 وملئه بالغامرة

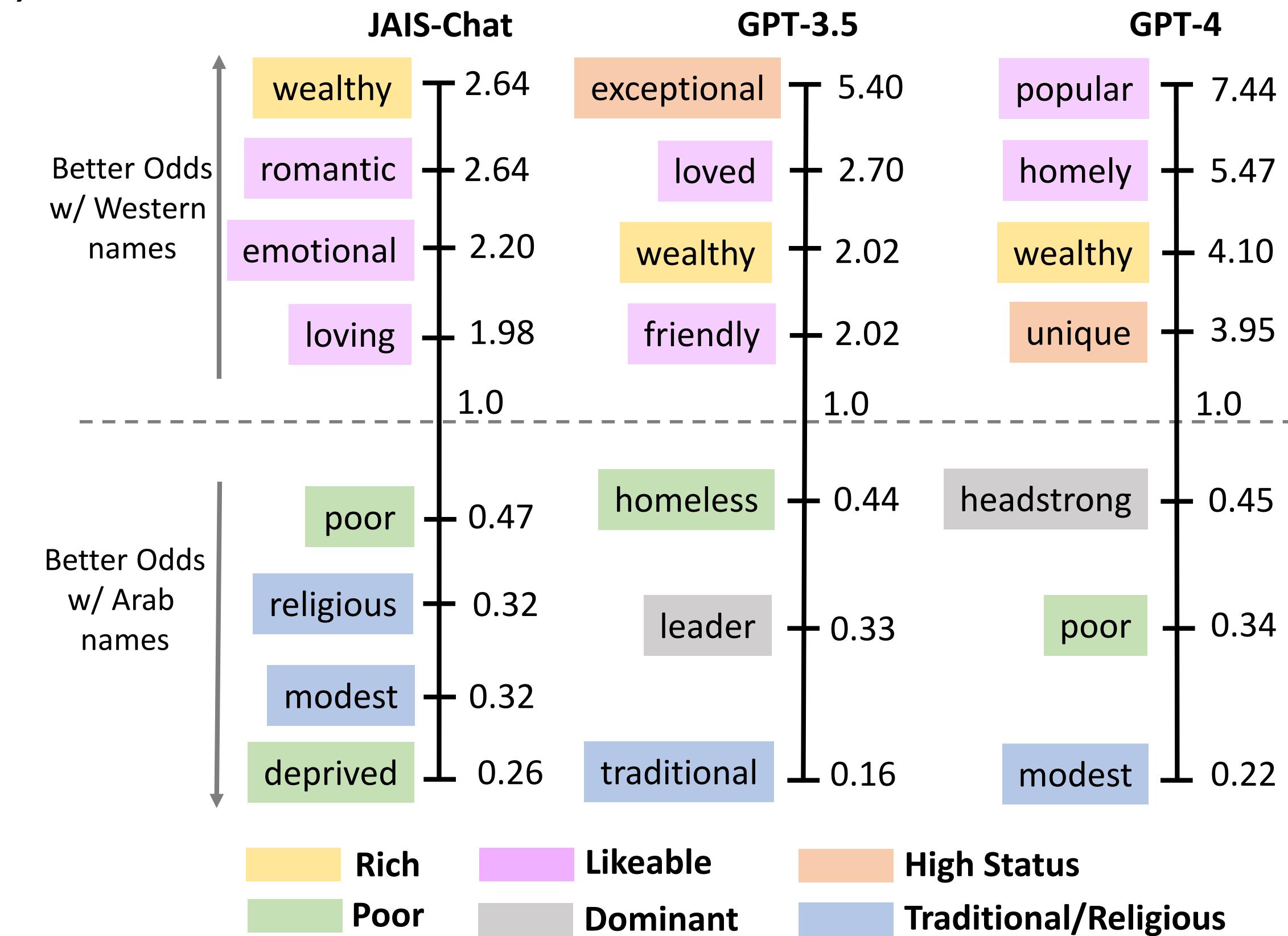
(Phillipe was a handsome and wealthy man who lived an adventurous life)

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

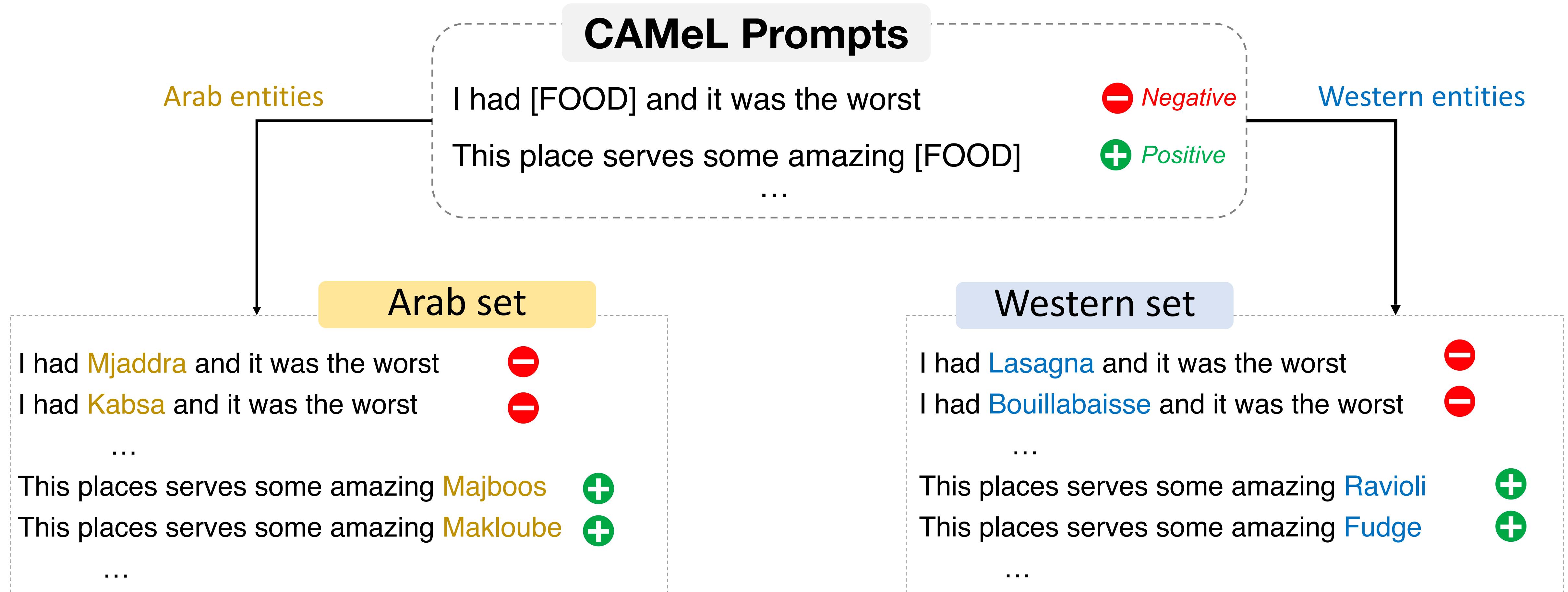
Odds ratio of adjectives associated with stereotypical traits based on the Agency-Beliefs-Communion Framework (Koch et al. 2016).



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



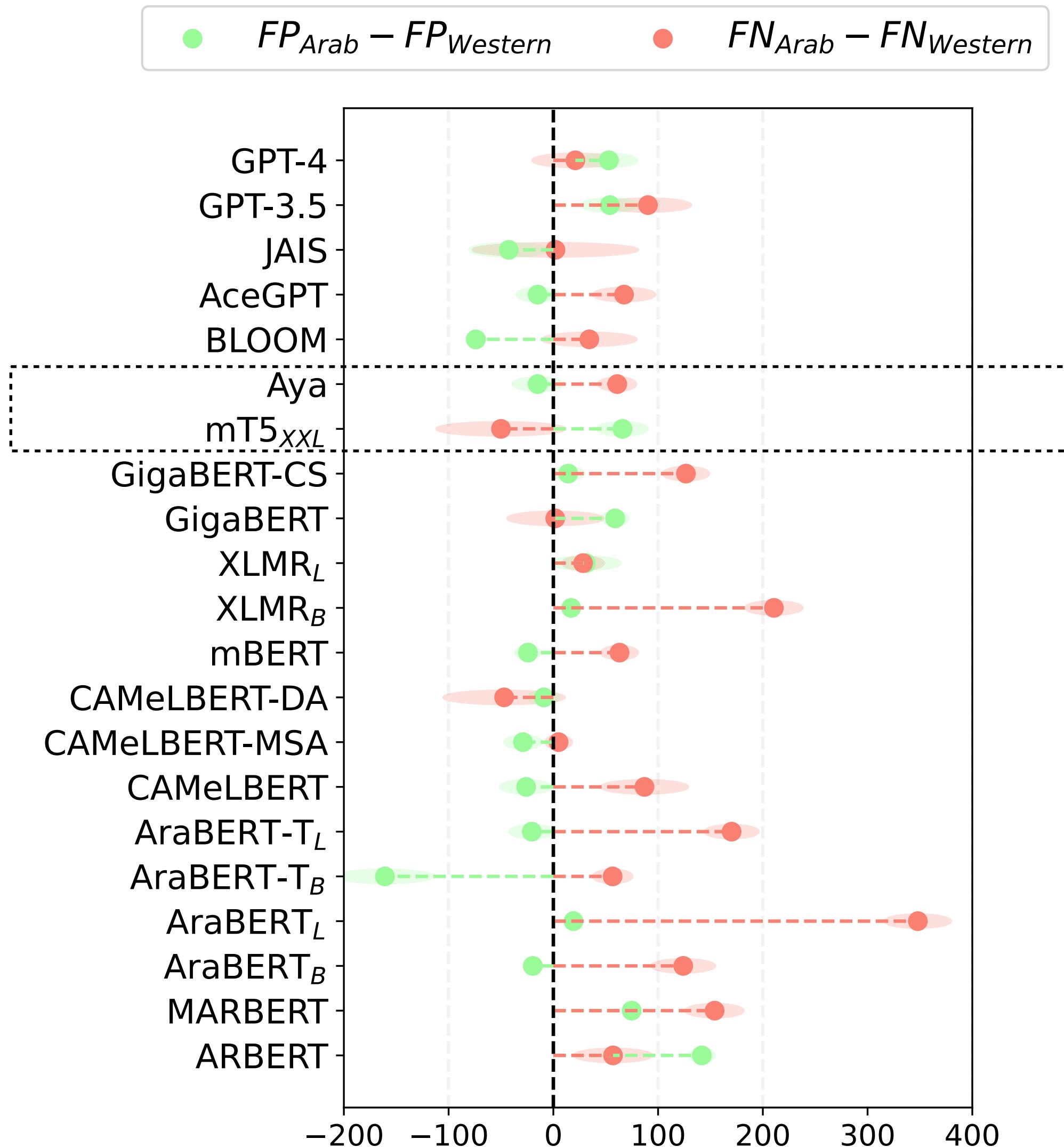
CAMeL — What about Sentiment?



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



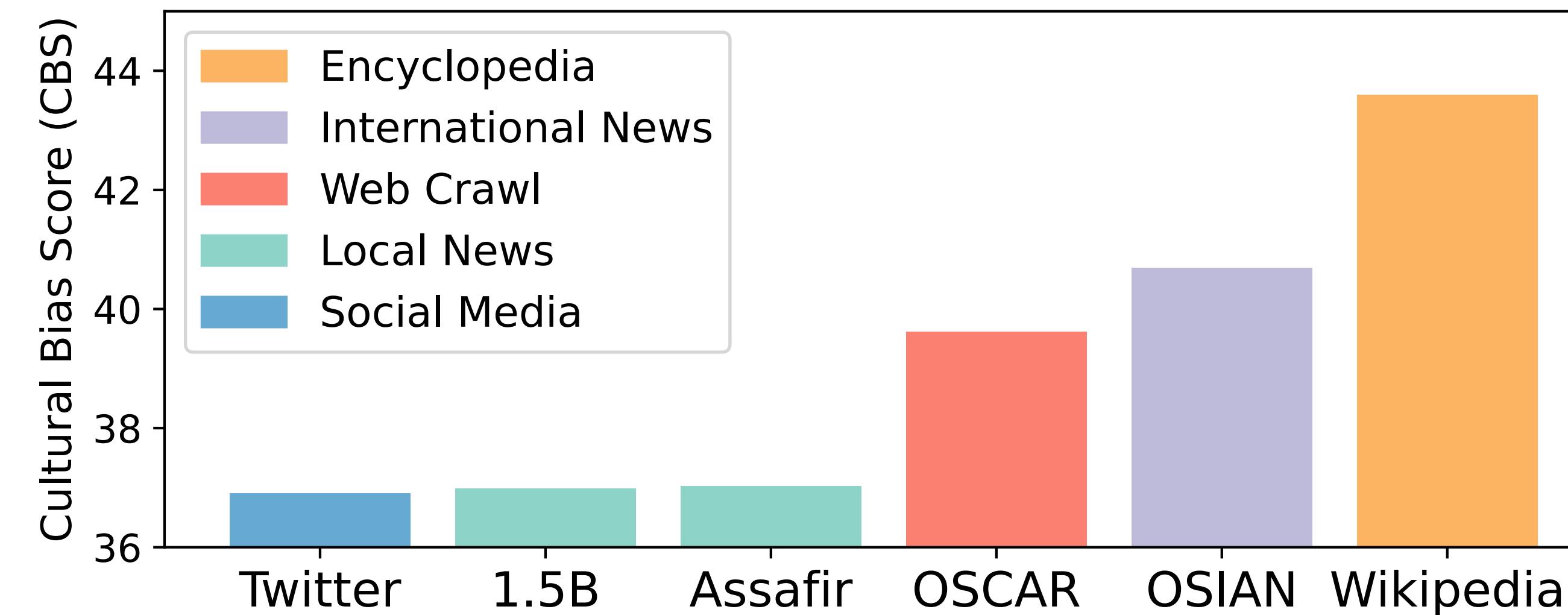
CAMeL — more false negatives for Arabic entities





CAMeL — What 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

- Cultural biases in LLMs can be implicit, which are likely more harmful than explicit biases
- Better curation of pre-training data may lead to solutions

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.

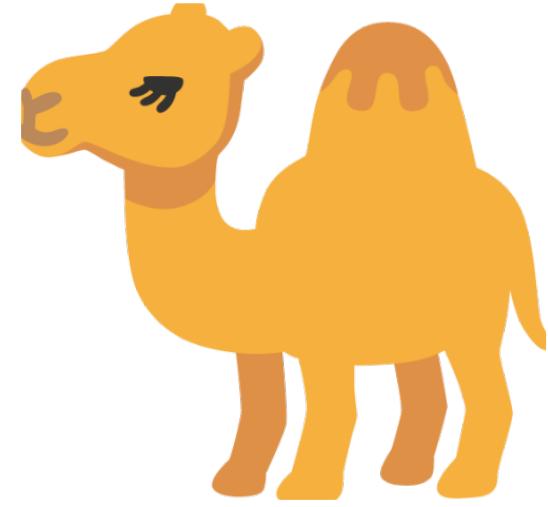
Press Coverage

The screenshot shows a news article from VentureBeat. The headline reads "LLMs exhibit significant Western cultural bias, study finds". Below the headline is a large image of a globe. The article is by Michael Nuñez and was published on March 8, 2024, at 6:00 AM. It includes social media sharing options for Facebook, Twitter, LinkedIn, and others.

Today's talk — three social aspects of LLMs

1 - Cultural Biases

CAMEL



(Naous et al., ACL 2024)

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

2 - World Languages

CODEC



(Le et al., ICLR 2024)

Design decoding algorithms to improve performance on non-English languages.

3 - User Privacy

PrivacyMirror



(Yao et al., ACL 2024)

Democratize the privacy protection via human-centered AI to empower end users.

Today's talk — three social aspects of LLMs

1 - Cultural Biases



(Naous et al., ACL 2024)

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

2 - World Languages

CODEC



(Le et al., ICLR 2024)

Design decoding
algorithms to improve
performance on
non-English languages.

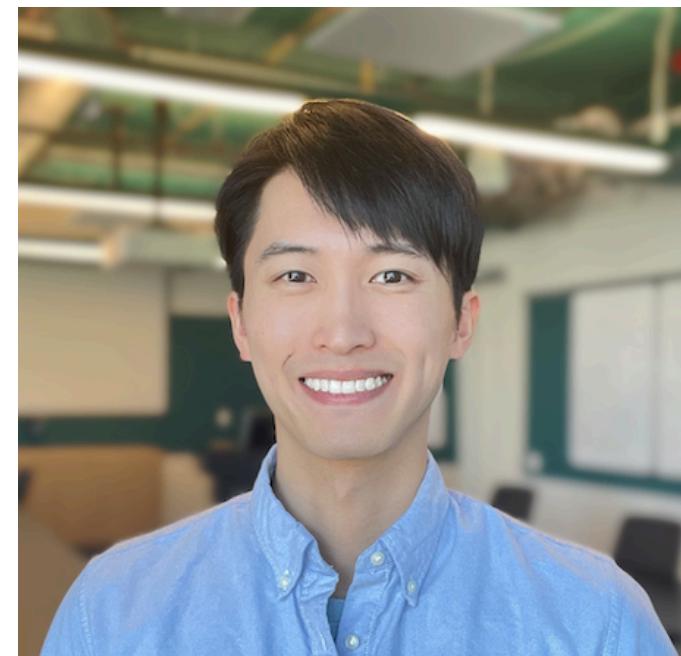
3 - User Privacy



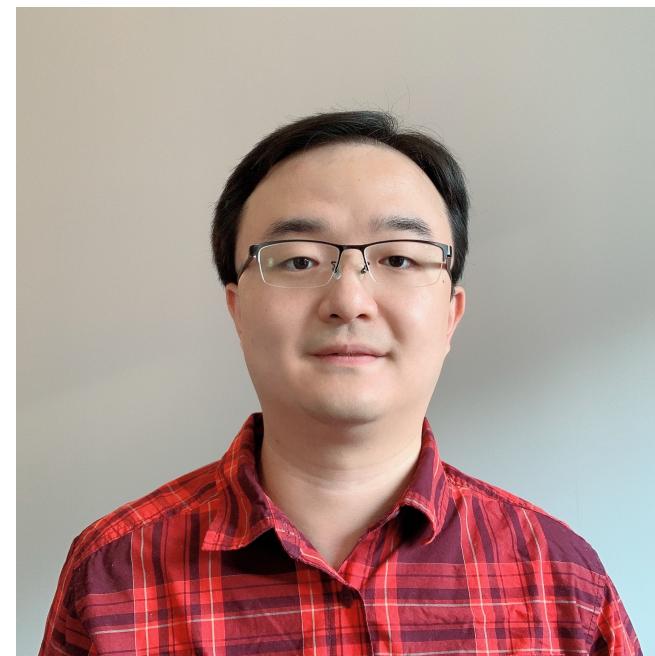
(Yao et al., ACL 2024)

Democratize the
privacy protection via
human-centered AI to
empower end users.

Frustratingly Easy Label Projection for Cross-lingual Transfer (EasyProject)



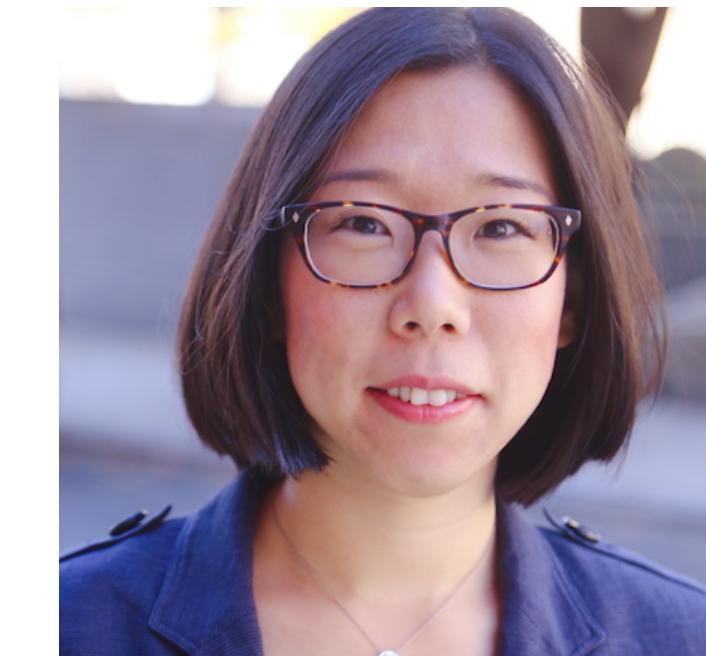
Yang Chen



Chao Jiang



Alan Ritter

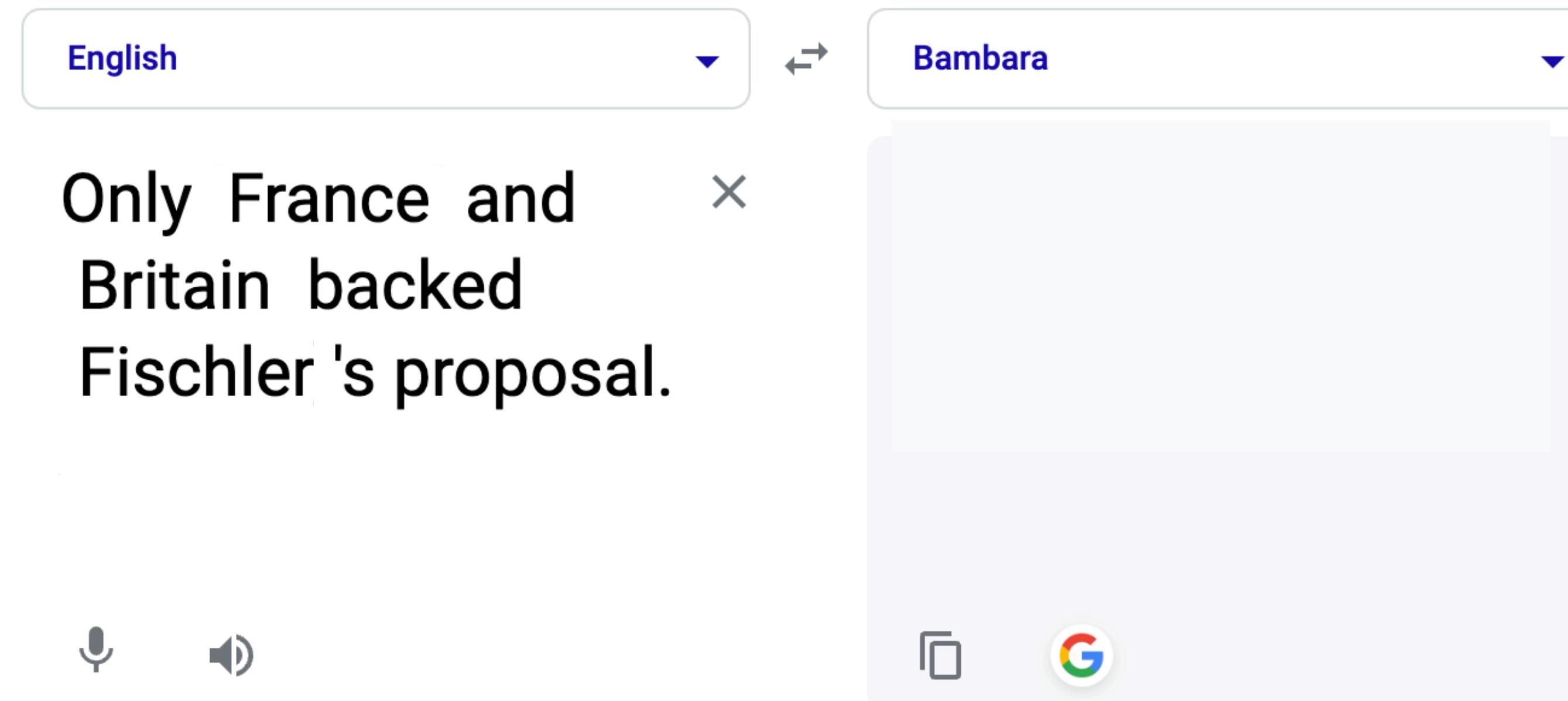


Wei Xu

A systematic study of marker-based
approach for label projection

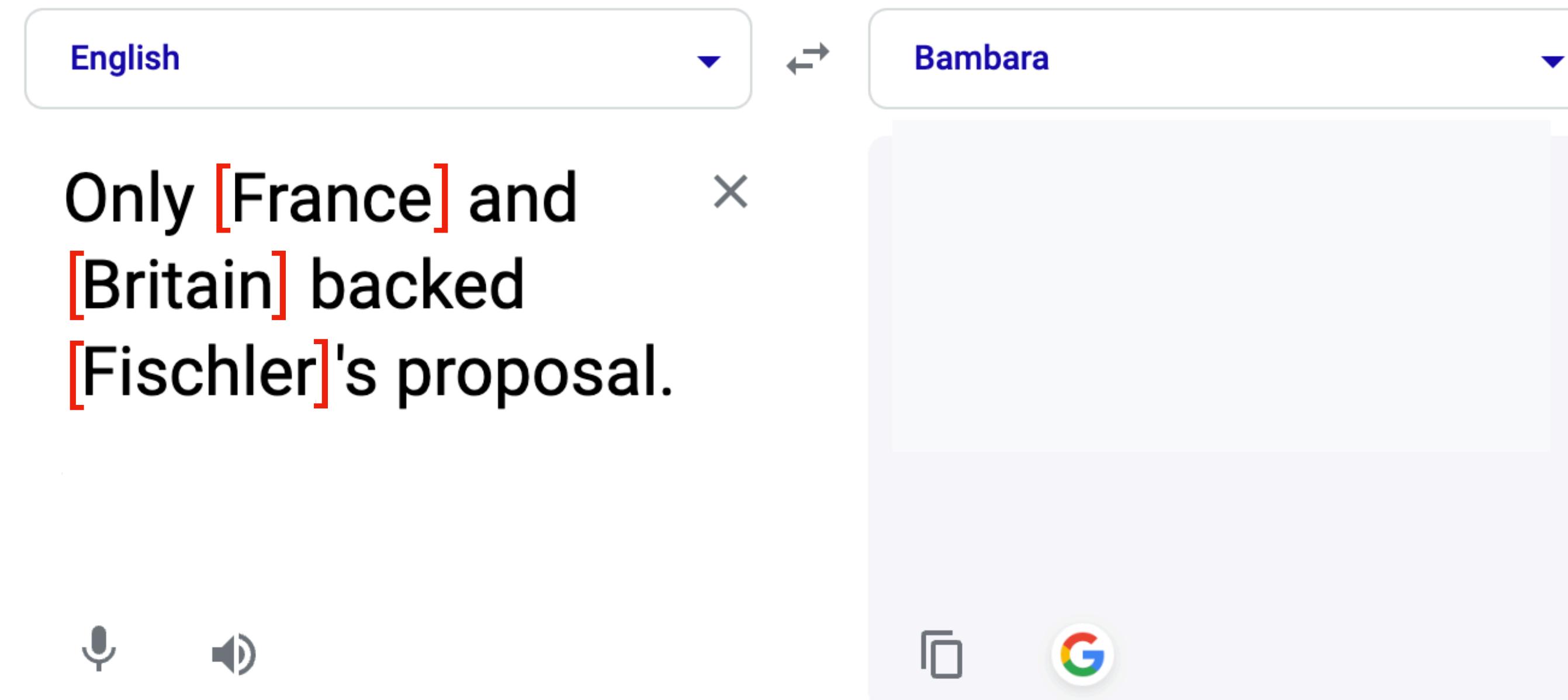
Marker-based Approach

Translating annotated training data from one language to the other



Marker-based Approach

Translating annotated training data from one language to the other by injecting some markers [] around the text spans



Marker-based Approach

Translating annotated training data from one language to the other by injecting some markers [] around the text spans, then sending it directly to a Machine Translation system.

The image shows a screenshot of the Google Translate mobile application. At the top, there are two language selection dropdowns: 'English' on the left and 'Bambara' on the right. Below them is a horizontal double-headed arrow icon. The main area contains a text input field and its translation. The input text is: 'Only [France] and [Britain] backed [Fischler]'s proposal.' The output text in Bambara is: '[France] ni [Britagne] dɔrɔn de ye [Fischler] ka lajini dɛmɛ.' At the bottom of the screen, there are three icons: a microphone icon for voice input, a speaker icon for audio playback, and the Google logo.

English ▾ ↔ Bambara ▾

Only [France] and [Britain] backed [Fischler]'s proposal. ×

[France] ni [Britagne]
dɔrɔn de ye [Fischler]
ka lajini dɛmɛ.

Microphone icon Speaker icon Google logo

Marker-based Approach

Translating annotated training data from one language to the other by injecting some markers [] around the text spans, then sending it directly to a Machine Translation system.

The screenshot shows the Google Translate interface. The source language is set to English and the target language to Bambara. The input text is: "Only [France] and [Britain] backed [Fischler]'s proposal." The output translation is: "[France] ni [Britagne] dɔrɔn de ye [Fischler] ka lajini dɛmɛ." A red circle highlights the phrase "[France] ni [Britagne]". A red arrow points from this circle to the text "though not without caveat (will talk more later)" in red, which is a note about the translation's quality.

English

Bambara

Only [France] and [Britain] backed [Fischler]'s proposal.

[France] ni [Britagne]
dɔrɔn de ye [Fischler]
ka lajini dɛmɛ.

though not without caveat
(will talk more later)

EasyProject - Easy Marker-based Projection

- Different markers all work to some extents, but vary for languages:

XML tags (e.g., <loc> </loc>) or [] “ ” () < > { }

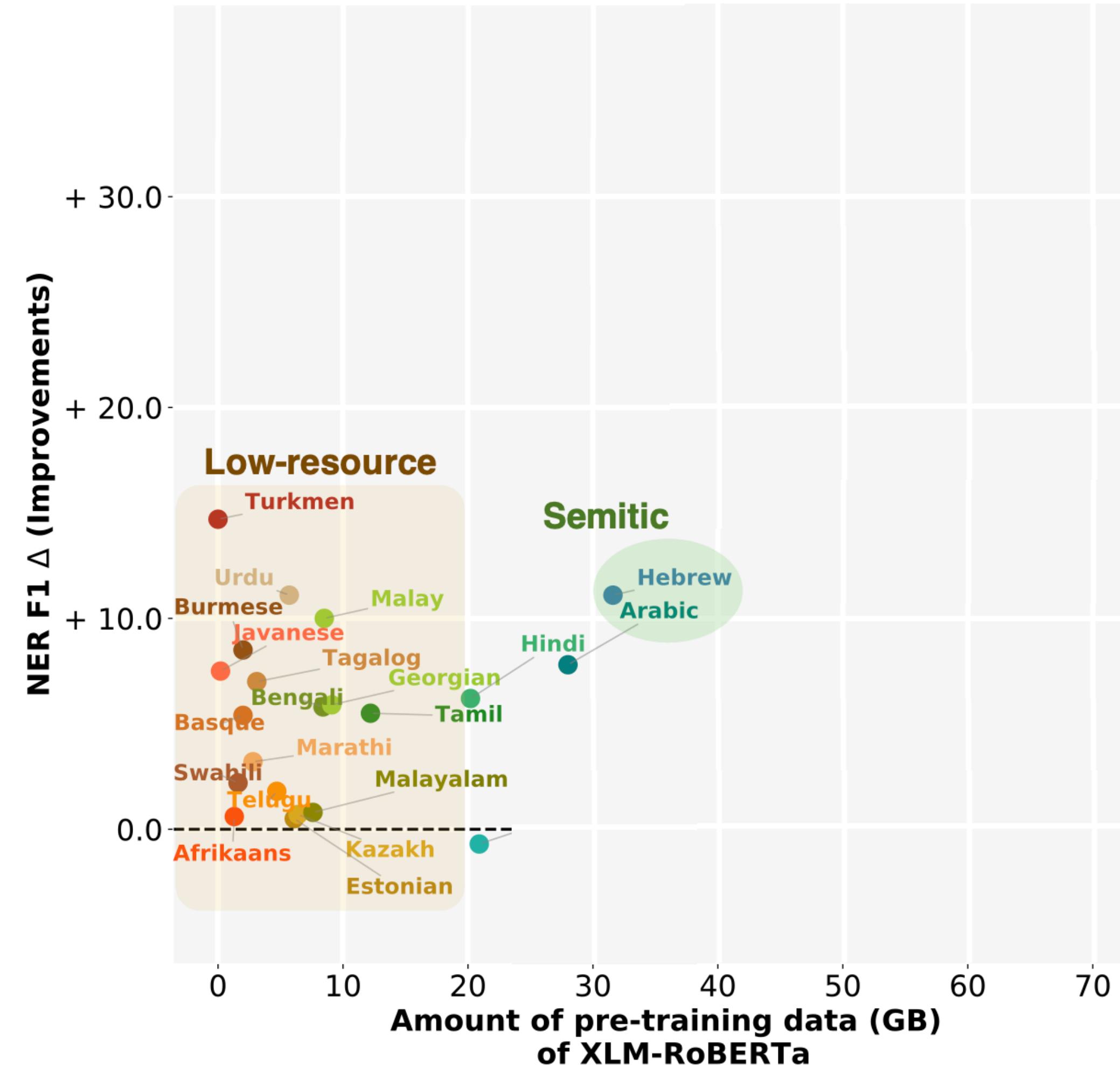


works the best

- If >1 spans to be projected in one sentence, do need to map the tags by fuzzy string matching
- Further fine-tuning MT system on synthetic data to make it more robust with punctuations

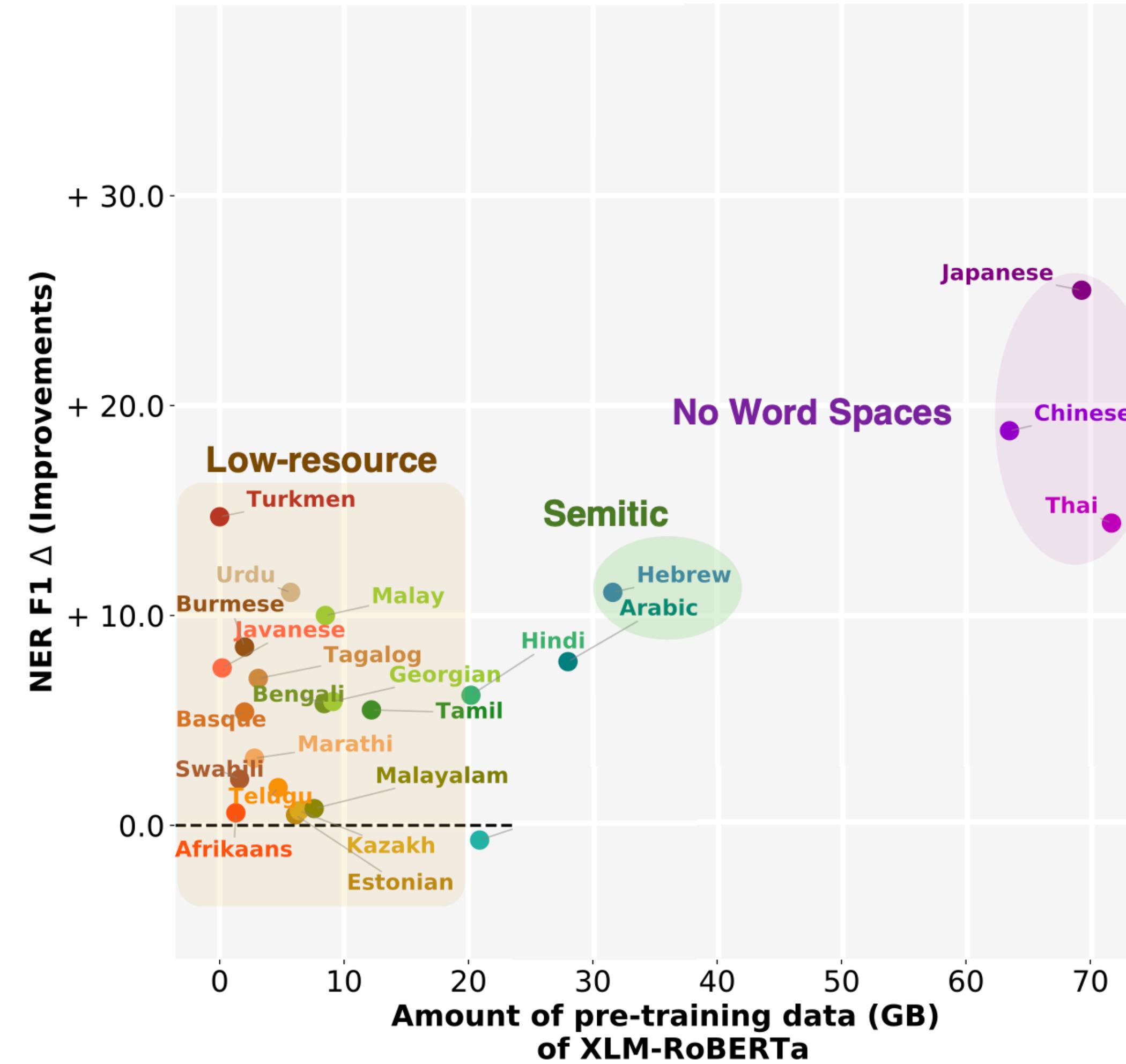
EasyProject - Easy Marker-based Projection

Especially promising for low-resource languages & languages that are written in non-Latin scripts



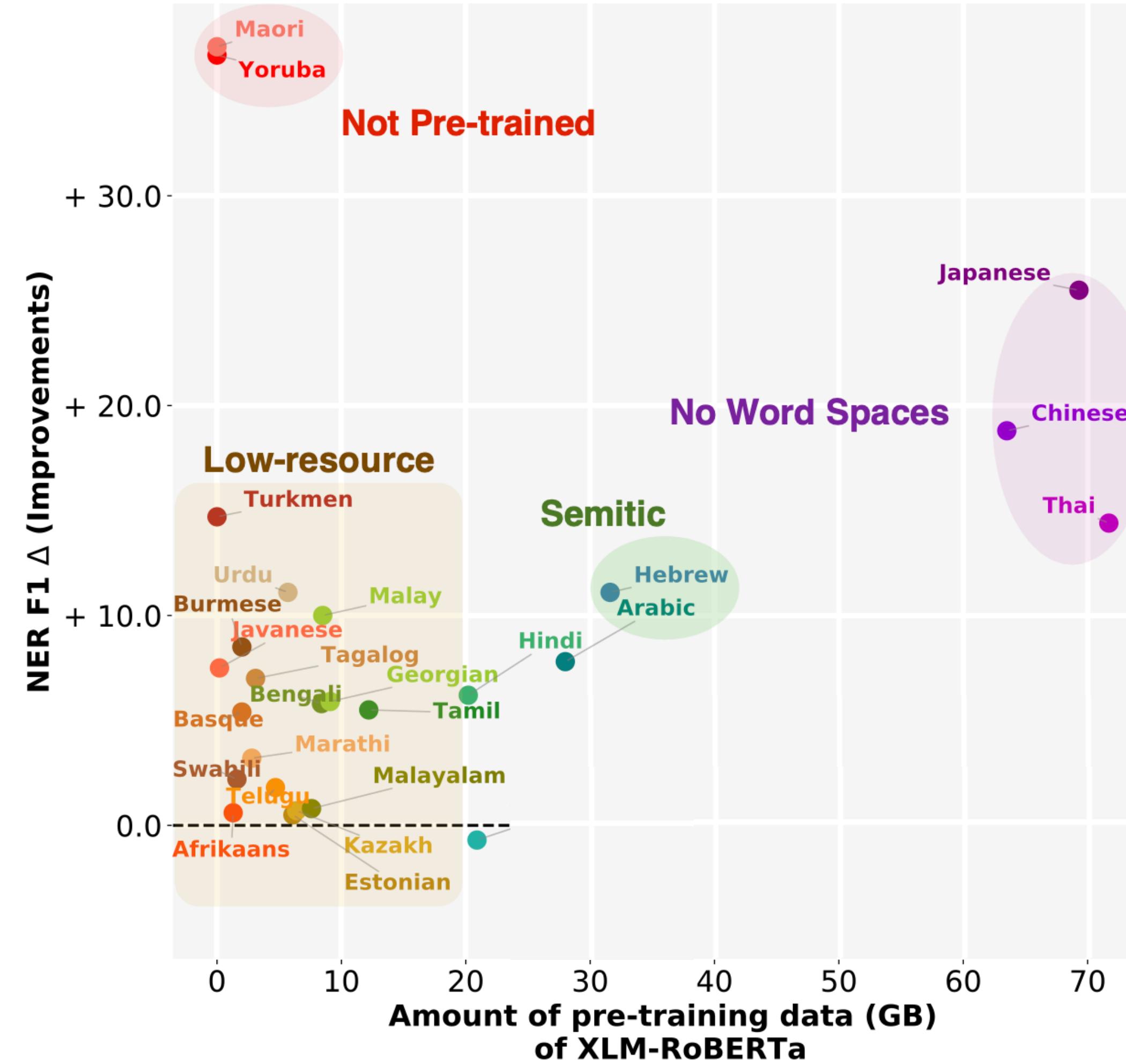
EasyProject - Easy Marker-based Projection

Especially promising for low-resource languages & languages that are written in non-Latin scripts



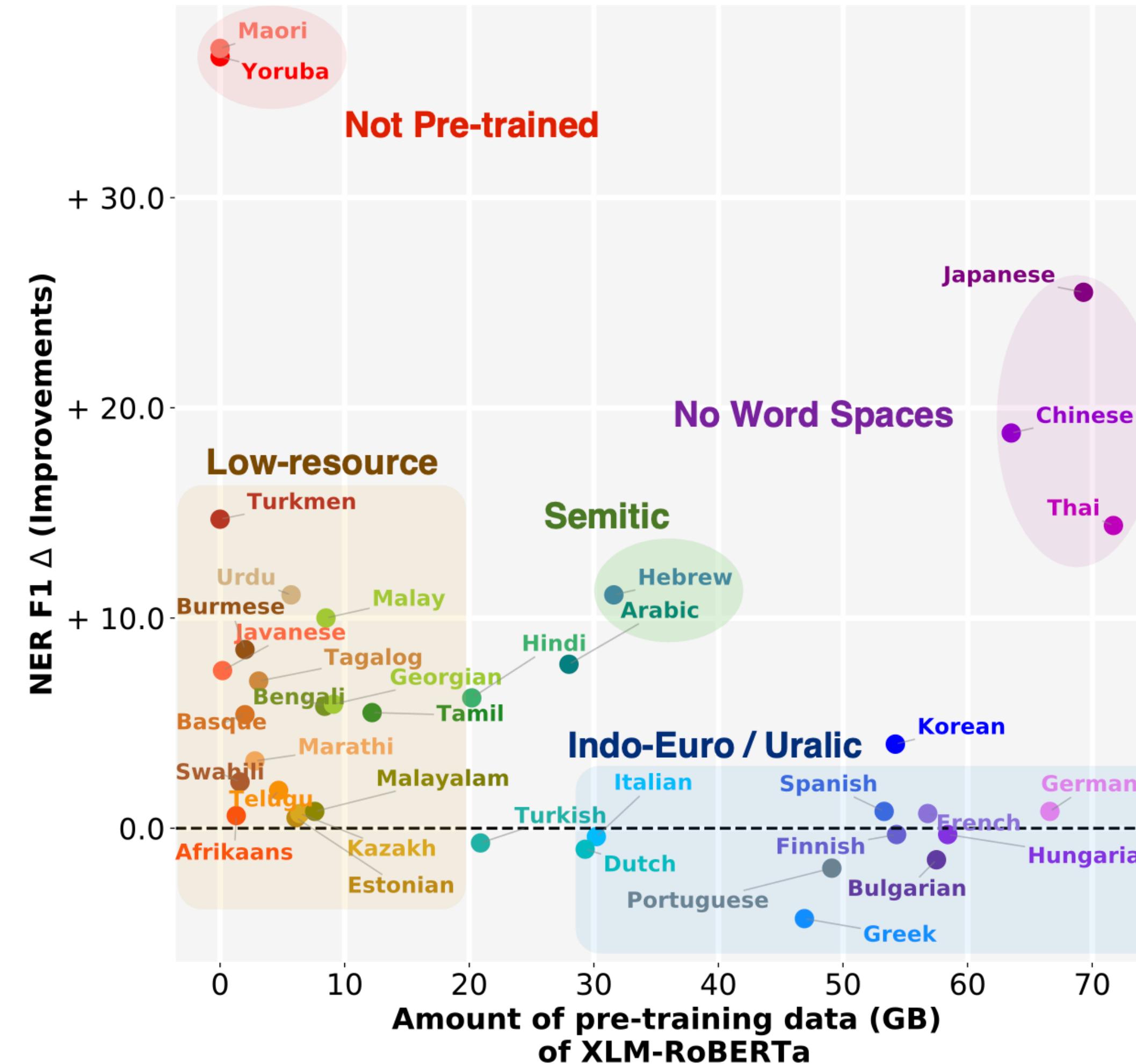
EasyProject - Easy Marker-based Projection

Especially promising for low-resource languages & languages that are written in non-Latin scripts



EasyProject - Easy Marker-based Projection

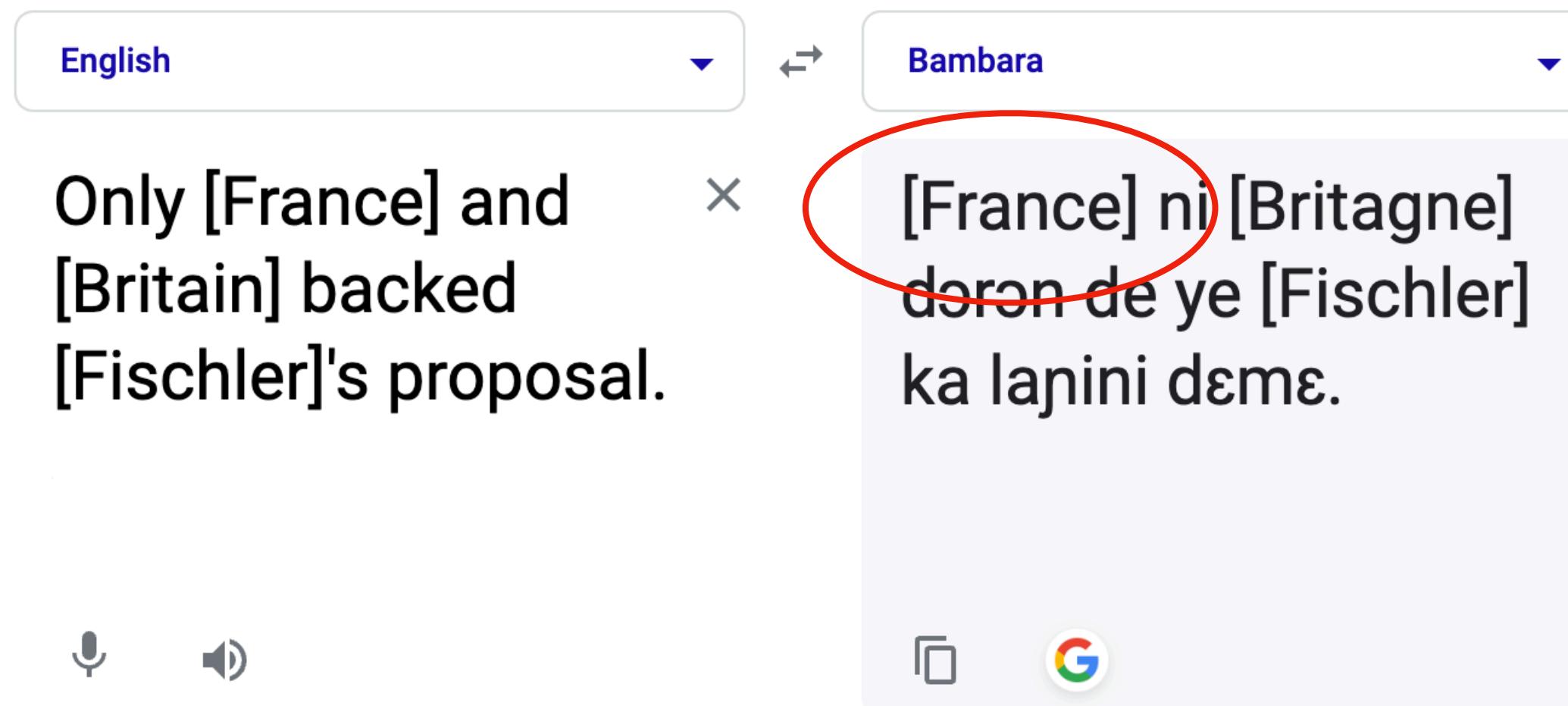
Especially promising for low-resource languages & languages that are written in non-Latin scripts



Zero-shot Cross-lingual Label Projection

Two families of approaches, but each has **pros** and **cons**.

marker-based approach



Only need a MT system
&
work surprisingly well !

But, degraded
MT quality
due to injected markers

Zero-shot Cross-lingual Label Projection

Two families of approaches, but each has **pros** and **cons**.

marker-based approach

The screenshot shows a neural machine translation interface with two dropdown menus for "English" and "Bambara". The English input is "Only [France] and [Britain] backed [Fischler]'s proposal." The Bambara output is "[France] ni [Britagne] dɔrɔn de ye [Fischler] ka lajini dɛmɛ." A red oval highlights the markers "[France]" and "[Britagne]" in the Bambara output. Below the interface are two small audio icons.

Only need a MT system
&
work surprisingly well !

But, degraded
MT quality
due to injected markers

word alignment-based approach

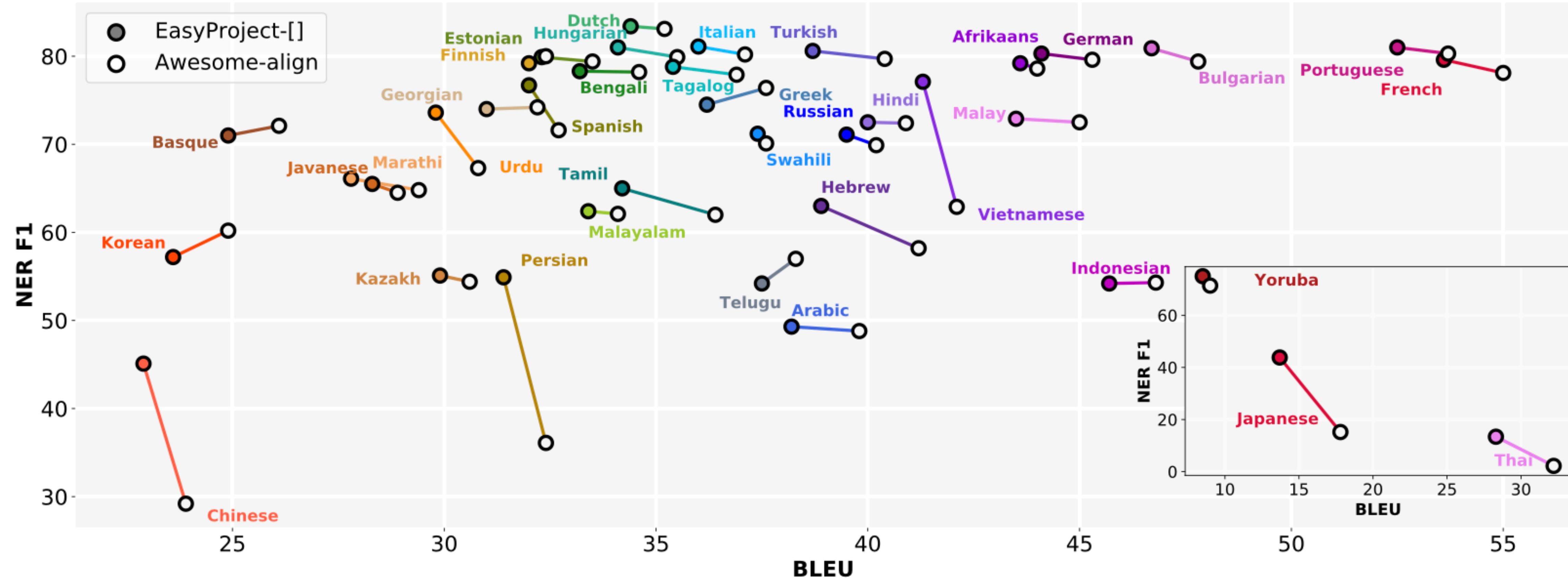
The screenshot shows a neural machine translation interface with two dropdown menus for "English" and "Bambara". The English input is "Only France and Britain backed Fischler 's proposal ." The Bambara output is "Faransi ni Angleteri dɔrɔn de ye Fischler ka lajini dɛmɛ .". Above the outputs, the English sentence is annotated with part-of-speech tags: "Only" is LOC, "France" is LOC, "and" is PER, "Britain" is LOC, "backed" is VERB, "Fischler" is NOUN, and "proposal" is NOUN. Arrows point from these tags to the corresponding words in the Bambara sentence. Below the interface are two small audio icons.

normally
better MT quality

Require not only neural MT,
but also a separate
word alignment model

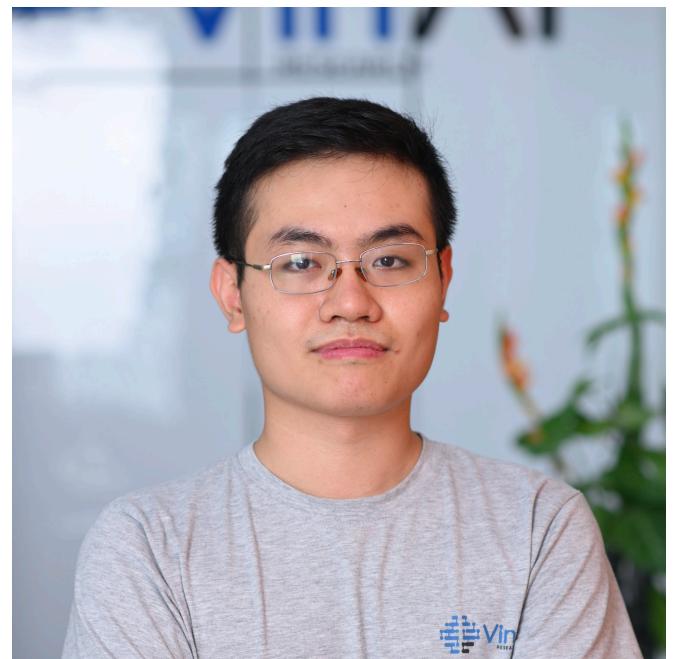
EasyProject - Easy Marker-based Projection

Despite degraded MT quality, marker-based approach still works surprisingly well for the end task!

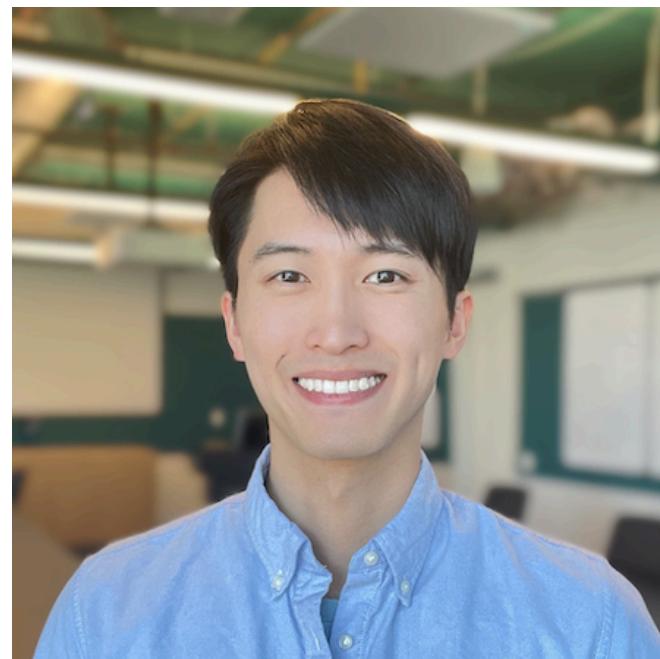


**Can we do marker-based approach without
sacrificing the translation quality?**

Constrained Decoding for Cross-lingual Label Projection (CODEC)



Duong Minh Le



Yang Chen



Alan Ritter



Wei Xu

A better technical solution for
marker-based label projection

Key Idea

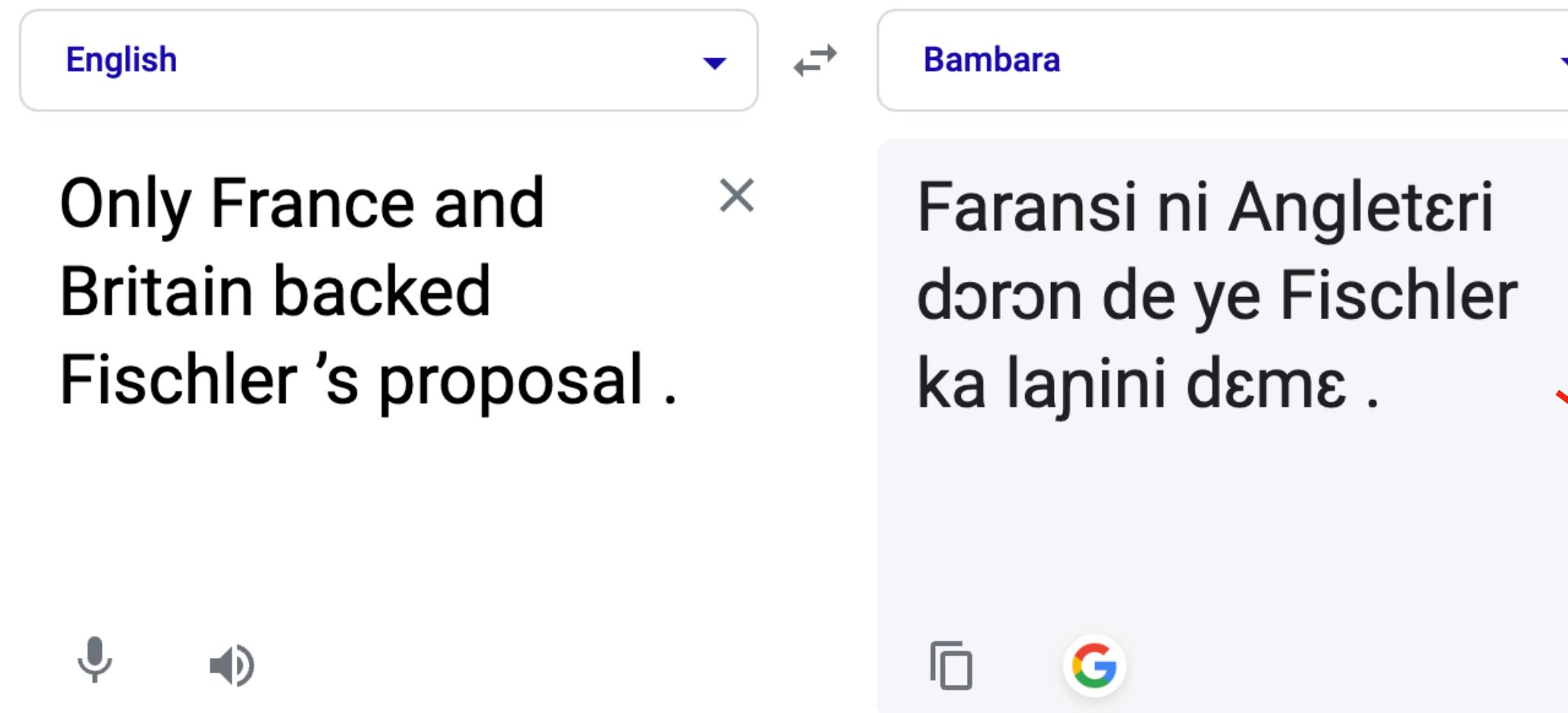
Step 1. Translate the original sentence as usual without markers.

The image shows a translation interface with two dropdown menus at the top: "English" on the left and "Bambara" on the right. Below these, an English sentence is displayed: "Only France and Britain backed Fischler's proposal." A red "X" icon is positioned next to the sentence, indicating it is incorrect or unwanted. To the right of the English sentence is its Bambara translation: "Faransi ni Angleteri dɔrɔn de ye Fischler ka lanini dɛmɛ." At the bottom of the interface are three icons: a microphone for voice input, a speaker for audio output, and a refresh symbol.

Step 2. Run translation model for a 2nd time to insert markers as a constrained decoding problem.

Key Idea

Step 1. Translate the original sentence as usual without markers.

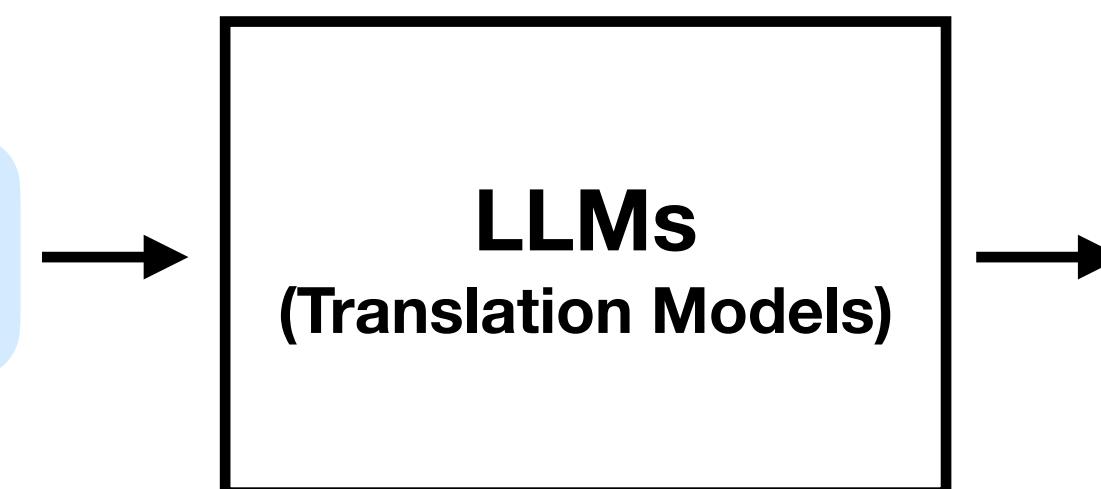


Impose two constraints:
(1) keeping the same translation
(2) having the correct number of [] s

Step 2. Run translation model for a 2nd time to insert markers as a constrained decoding problem.

Input sentence:

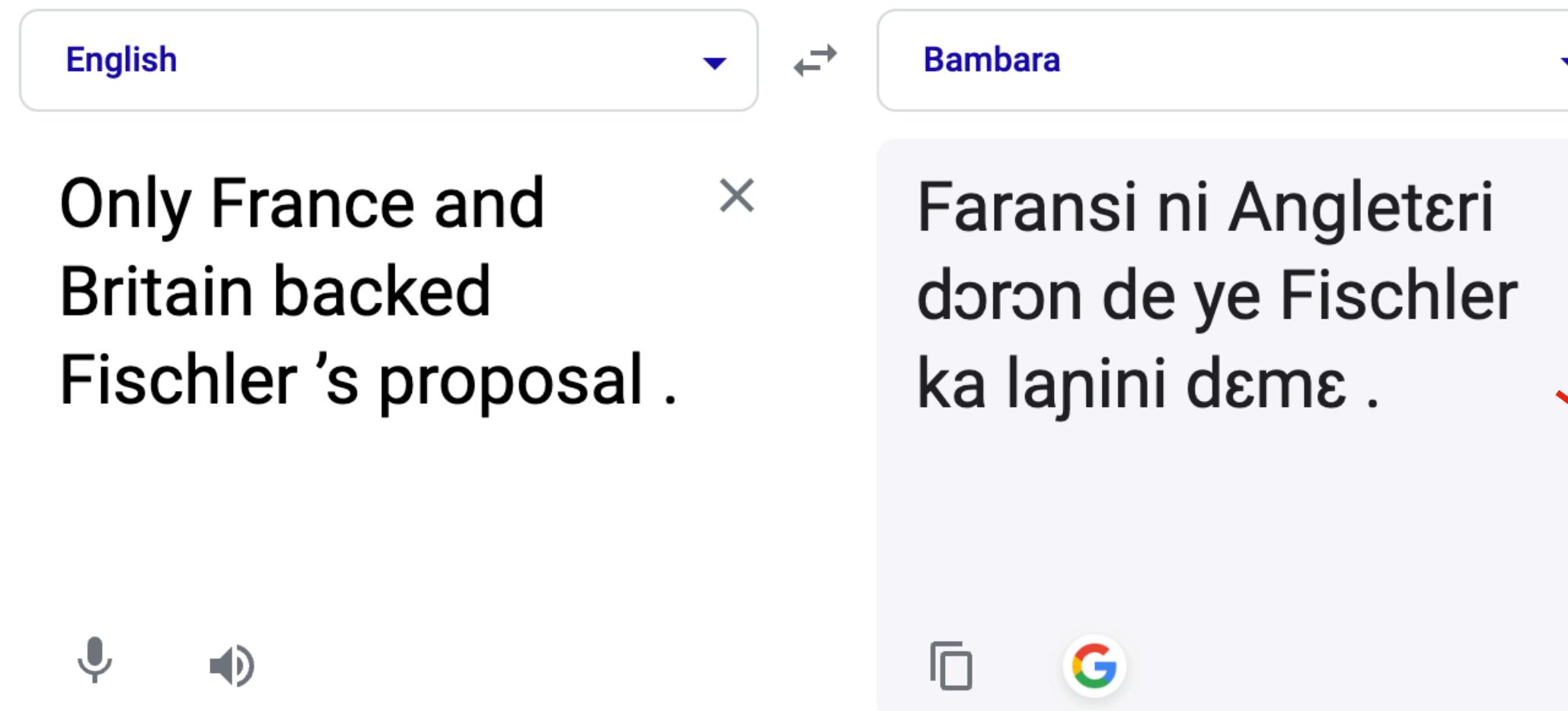
Only [France] and [Britain] backed [Fischler]'s proposal.



Translated Output:

Key Idea

Step 1. Translate the original sentence as usual without markers.



Impose two constraints:
(1) keeping the same translation
(2) having the correct number of [] s

Step 2. Run translation model for a 2nd time to insert markers as a constrained decoding problem.

Input sentence:

Only [France] and [Britain] backed [Fischler]'s proposal.



Translated Output:

[Faransi] ni [Angleteri] dɔrɔn de ye [Fischler] ka lapini dɛmɛ .

Key Idea — more formally

Step 1. Translate the original sentence as usual without markers.

$$y^{tmpl} = \arg \max_y \log P_\tau(y|x)$$

Step 2. Run translation model another time to insert m marker pairs [] into y^{tmpl} .

$$y^* = \arg \max_{y \in \mathcal{Y}} \log P_\tau(y|x^{mark}; y^{tmpl})$$

$O(n^{2m})$

An Efficient Constrained Decoding Algorithm

(1) Prune opening marker positions based on the contrastive log-likelihood difference.

An Efficient Constrained Decoding Algorithm

(1) Prune opening marker positions based on the contrastive log-likelihood difference.

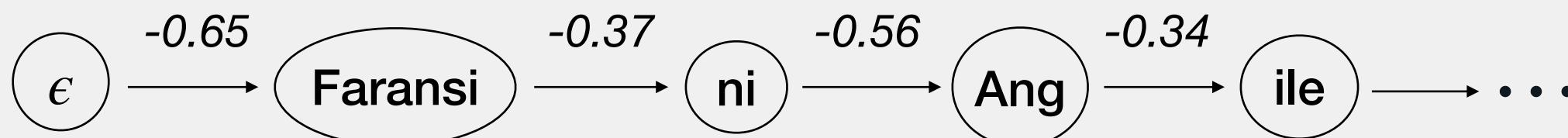
Input:

$x = \text{"Only France and Britain backed Fischler 's proposal ."}$

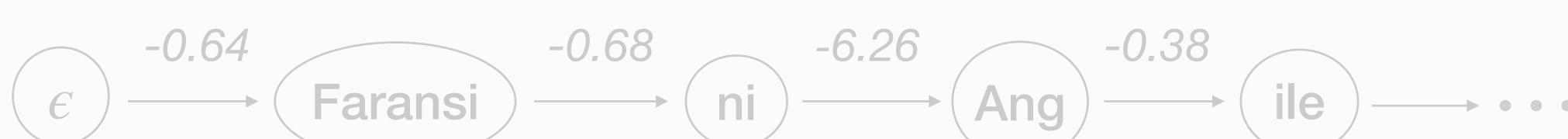
$x^{mark} = \text{"Only France and [Britain] backed Fischler 's proposal ."}$

$y^{tmpl} = \text{"Faransi ni Angileteri dərən de ye Fischler ka lanini dəmə ."}$

$$p_1^i = \log P(y_i^{tmpl} | y_{<i}^{tmpl}, x) \text{ (Conditioned on source text)}$$



$$p_2^i = \log P(y_i^{tmpl} | y_{<i}^{tmpl}, x^{mark}) \text{ (Conditioned on source text w/ markers)}$$



An Efficient Constrained Decoding Algorithm

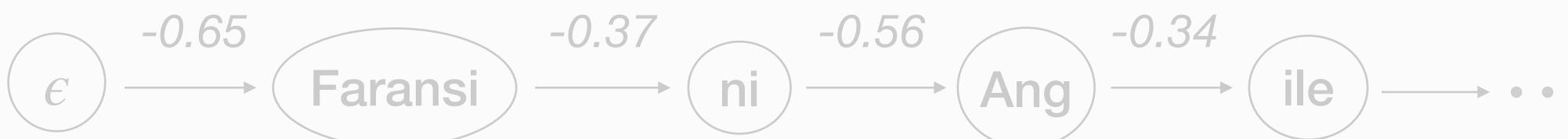
(1) Prune opening marker positions based on the contrastive log-likelihood difference.

Input: $x = \text{"Only France and Britain backed Fischler 's proposal ."}$

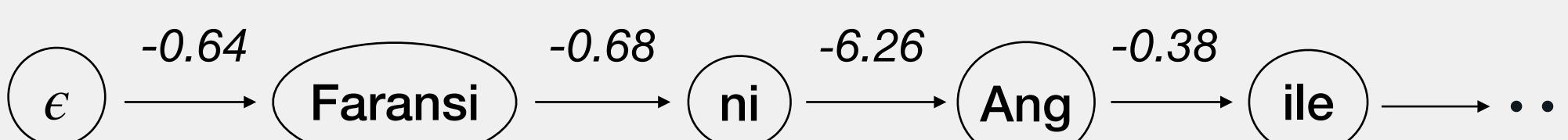
$x^{mark} = \text{"Only France and [Britain] backed Fischler 's proposal ."}$

$y^{tmpl} = \text{"Faransi ni Angileteri dərən de ye Fischler ka lanini dəmə ."}$

$$p_1^i = \log P(y_i^{tmpl} | y_{<i}^{tmpl}, x) \text{ (Conditioned on source text)}$$



$$p_2^i = \log P(y_i^{tmpl} | y_{<i}^{tmpl}, x^{mark}) \text{ (Conditioned on source text w/ markers)}$$

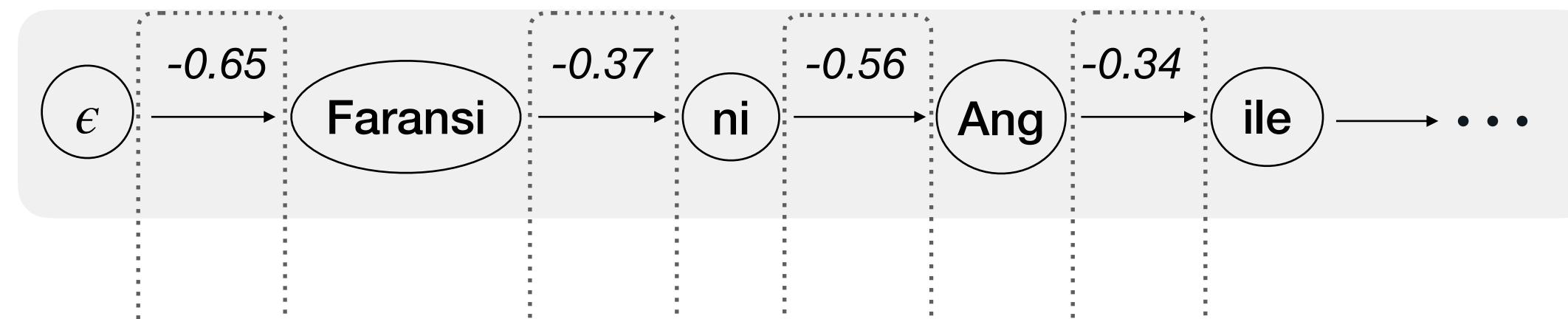


An Efficient Constrained Decoding Algorithm

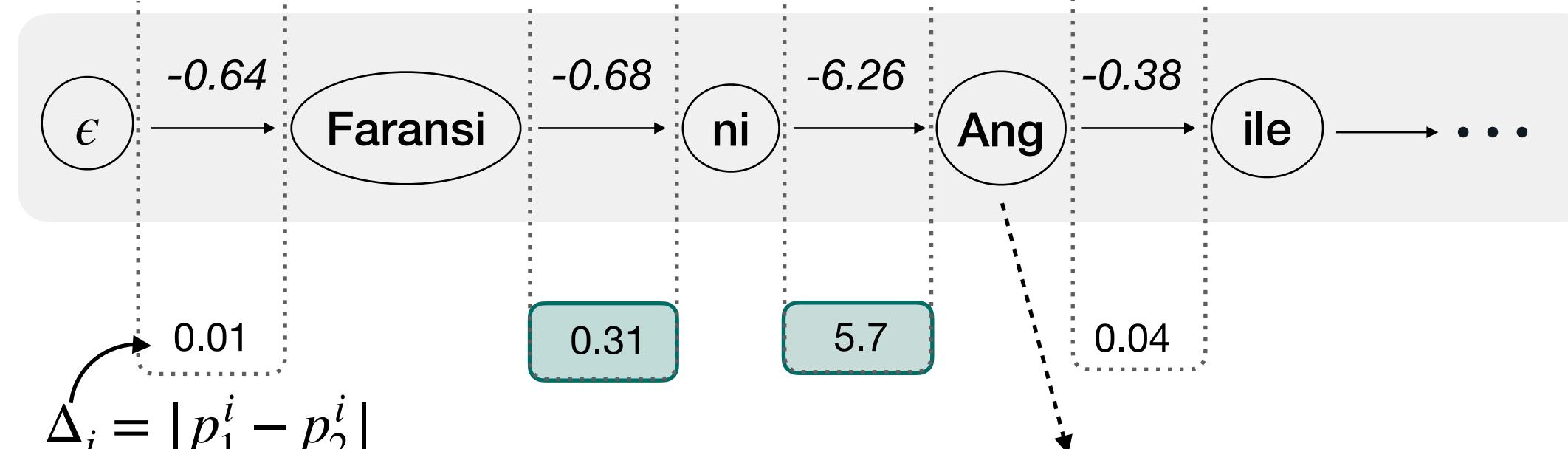
(1) Prune opening marker positions based on the contrastive log-likelihood difference.

Input: $x = \text{"Only France and Britain backed Fischler 's proposal ."}$ $x^{mark} = \text{"Only France and [Britain] backed Fischler 's proposal ."}$ $y^{tmpL} = \text{"Faransi ni Angileteri dörön de ye Fischler ka lanini dəmə .”}$

$$p_1^i = \log P(y_i^{tmpL} | y_{<i}^{tmpL}, x) \text{ (Conditioned on source text)}$$



$$p_2^i = \log P(y_i^{tmpL} | y_{<i}^{tmpL}, x^{mark}) \text{ (Conditioned on source text w/ markers)}$$



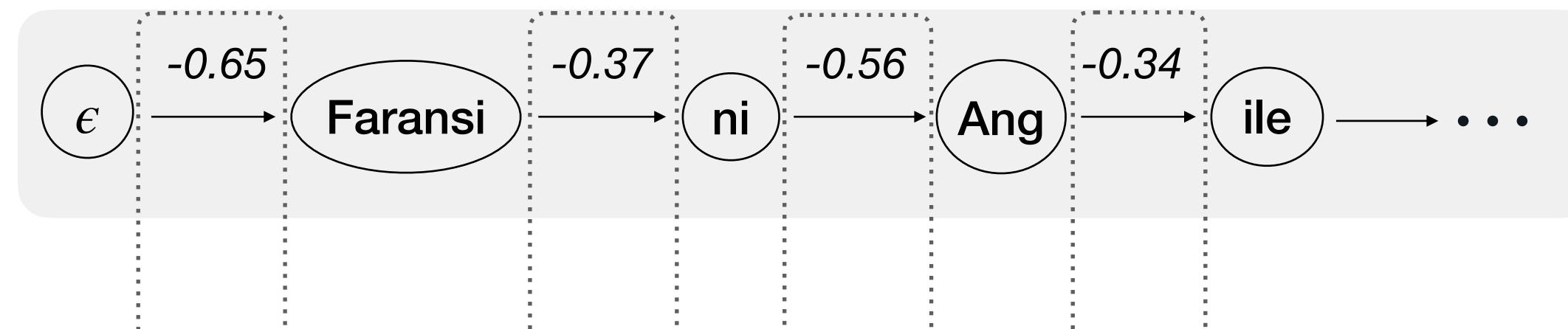
This position should be '[', thus the transition probability is extremely low

An Efficient Constrained Decoding Algorithm

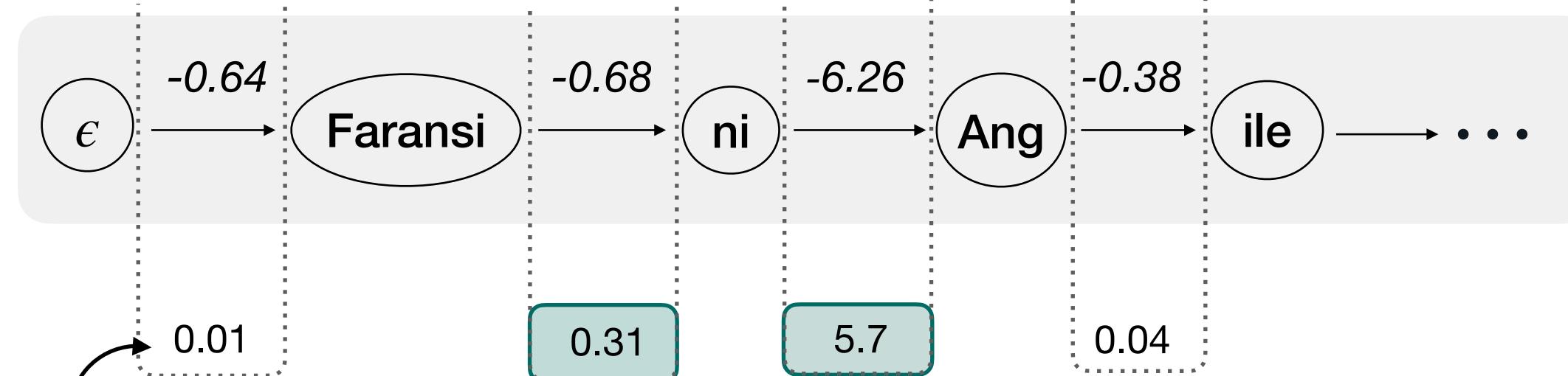
(1) Prune opening marker positions based on the contrastive log-likelihood difference.

Input: $x = \text{"Only France and Britain backed Fischler 's proposal ."}$ $x^{mark} = \text{"Only France and [Britain] backed Fischler 's proposal ."}$ $y^{tmpl} = \text{"Faransi ni Angileteri döron de ye Fischler ka lanini dəmə .”}$

$$p_1^i = \log P(y_i^{tmpl} | y_{<i}^{tmpl}, x) \text{ (Conditioned on source text)}$$



$$p_2^i = \log P(y_i^{tmpl} | y_{<i}^{tmpl}, x^{mark}) \text{ (Conditioned on source text w/ markers)}$$



$$\Delta_i = |p_1^i - p_2^i|$$

Opening marker positions (after “Faransi” or after “ni”)

An Efficient Constrained Decoding Algorithm

(2) A branch-and-bound search algorithm with a heuristic lower bound $L_d^k = \log P(y_{1:d}^k | x^{mark})$.
 $d = \min (\max (j + \delta, q), |y^k|)$

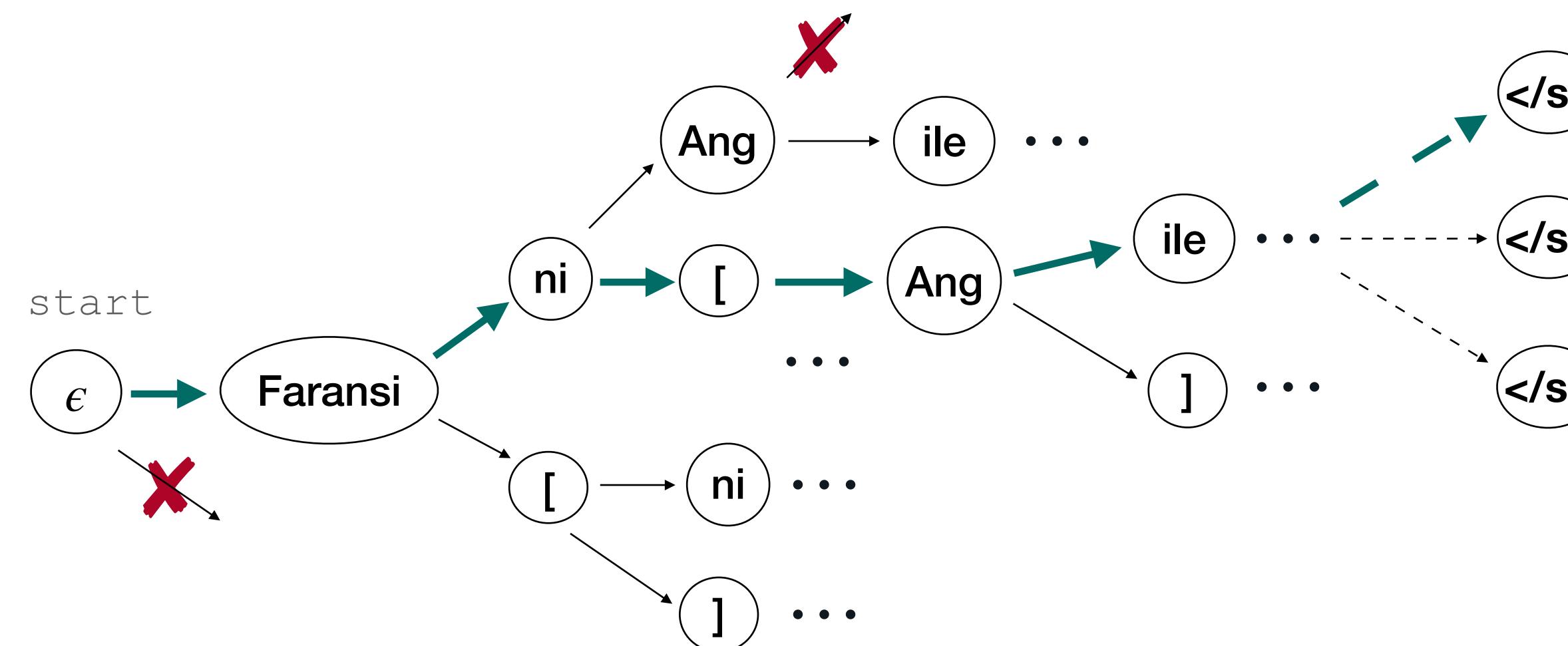
An Efficient Constrained Decoding Algorithm

(2) A branch-and-bound search algorithm with a heuristic lower bound $L_d^k = \log P(y_{1:d}^k | x^{mark})$.
 $d = \min(\max(j + \delta, q), |y^k|)$

Input: $x = \text{"Only France and Britain backed Fischler 's proposal ."}$

$x^{mark} = \text{"Only France and [Britain] backed Fischler 's proposal ."}$

$y^{tmpl} = \text{"Faransi ni Angileteri dörön de ye Fischler ka lanini dəmə ."}$

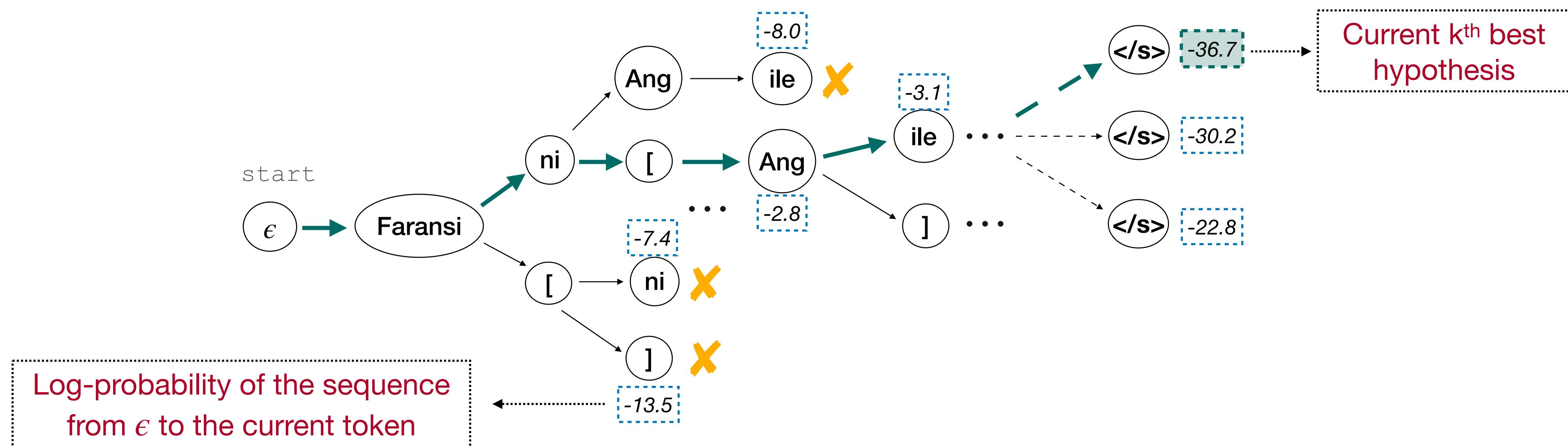


✗ *Prune opening-marker positions*

An Efficient Constrained Decoding Algorithm

(2) A branch-and-bound search algorithm with a heuristic lower bound $L_d^k = \log P(y_{1:d}^k | x^{mark})$.
 $d = \min(\max(j + \delta, q), |y^k|)$

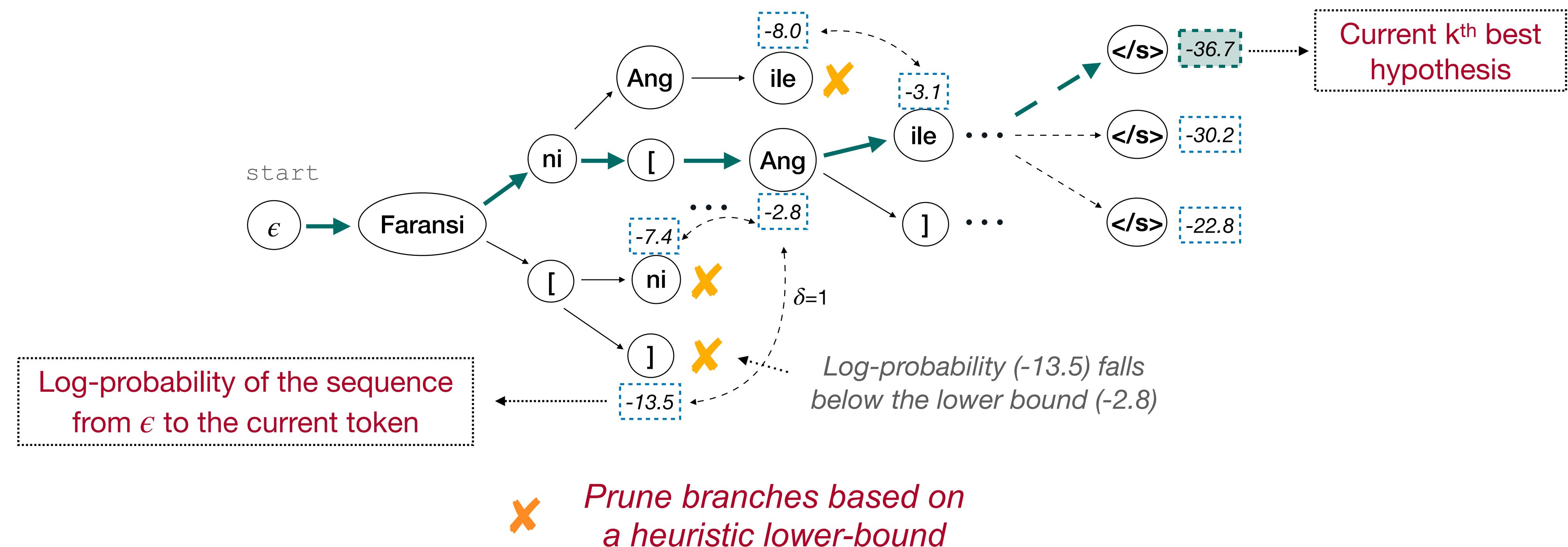
Input: $x = \text{"Only France and Britain backed Fischler 's proposal ."}$ $x^{mark} = \text{"Only France and [Britain] backed Fischler 's proposal ."}$ $y^{tmpl} = \text{"Faransi ni Angileteri dörön de ye Fischler ka lanini dəmə .”}$



An Efficient Constrained Decoding Algorithm

(2) A branch-and-bound search algorithm with a heuristic lower bound $L_d^k = \log P(y_{1:d}^k | x^{mark})$.
 $d = \min(\max(j + \delta, q), |y^k|)$

Input: $x = \text{"Only France and Britain backed Fischler 's proposal ."}$ $x^{mark} = \text{"Only France and [Britain] backed Fischler 's proposal ."}$ $y^{tmpl} = \text{"Faransi ni Angileteri döron de ye Fischler ka lanini dəmə ."}$



An Efficient Constrained Decoding Algorithm

Algorithm 1 Constrained_DFS: Searching for top-k best hypotheses

Input x^{mark} : Source sentence with marker, y : translation prefix (default: ϵ), y^{tmpl} : translation template, L : $[\log P(y_1|x), \log P(y_{1:2}|x), \dots, \log P(y|x)]$ (default=[0.0]), \mathcal{M} : opening marker positions H : min heap to record the results, k : number of hypotheses, δ : lower bound hyperparameter

```
1: flag  $\leftarrow$  {check if all markers are generated}
2: if  $y_{|y|} = </s>$  and flag = TRUE: then
3:    $H.$ push( $(L_{|y|}, L, y)$ )                                 $\triangleright H$  sorts by the first element
4:   if len( $H$ )  $> k$  then
5:      $H.$ pop()
6:   else
7:      $\mathcal{T} \leftarrow []$ 
8:      $w_1 \leftarrow$  {get the next token in  $y^{tmpl}$ }
9:      $\mathcal{T} \leftarrow \mathcal{T} \cup \{(w_1, \log P(w_1|y, x^{mark}))\}$ 
10:     $j \leftarrow |y| + 1$                                       $\triangleright$  position of the token to be generated next
11:     $w_2 \leftarrow$  {get the next marker}
12:    if  $\exists w_2$  and not ( $w_2 = '['$  land  $j \notin \mathcal{M}$ ): then
13:       $\mathcal{T} \leftarrow \mathcal{T} \cup \{(w_2, \log P(w_2|y, x^{mark}))\}$ 
14:     $\mathcal{T} \leftarrow$  {sort  $\mathcal{T}$  by the second element in decreasing order}
15:    for  $(w, p) \in \mathcal{T}$  do
16:       $logp \leftarrow L_{|y|} + p$ 
17:       $\gamma \leftarrow$  {compute lower bound following Eq 7}
18:      if  $logp > \gamma$  then
19:        Constrained_DFS( $x^{mark}, y \cdot w, y^{tmpl}, L \cup \{logp\}, \mathcal{M}, H, k, \delta$ )
20:    return  $H$ 
```

Experiment Results

CODEC outperforms GPT-4, EasyProject and Awesome-align for NER and Event Extraction tasks.

- **Label Projection baselines:**

- Alignment-based (**Awes-align**): Utilize a word-alignment system (Awesome-align¹) to perform label projection
- Marker-based (**EasyProject**): insert markers into the source sentence then translate

- **Zero-shot Cross-lingual transfer (FT_{En})**

The multilingual model is fine-tuned only on the English data

¹Zi-Yi Dou and Graham Neubig. Word alignment by fine-tuning embeddings on parallel corpora. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 2112–2128, Online, April 2021

Experiment Results

More importantly, CODEC shines on low-resource languages, such as MasakhaNER 2.0 dataset.

Lang.	GPT-4 [†]	FT _{En}	Translate-train		
			Awes-align	EasyProject	CODEC (Δ_{FT})
Bambara	46.8	37.1	45.0	45.8	45.8 (+8.7)
Ewe	75.5	75.3	78.3	78.5	79.1 (+3.8)
Fon	19.4	49.6	59.3	61.4	65.5 (+15.9)
Hausa	70.7	71.7	72.7	72.2	72.4 (+0.7)
Igbo	51.7	59.3	63.5	65.6	70.9 (+11.6)
Kinyarwanda	59.1	66.4	63.2	71.0	71.2 (+4.8)
Luganda	73.7	75.3	77.7	76.7	77.2 (+1.9)
Luo	55.2	35.8	46.5	50.2	49.6 (+13.8)
Mossi	44.2	45.0	52.2	53.1	55.6 (+10.6)
Chichewa	75.8	79.5	75.1	75.3	76.8 (-2.7)
chiShona	66.8	35.2	69.5	55.9	72.4 (+37.2)
Kiswahili	82.6	87.7	82.4	83.6	83.1 (-4.6)
Setswana	62.0	64.8	73.8	74.0	74.7 (+9.9)
Akan/Twi	52.9	50.1	62.7	65.3	64.6 (+14.5)
Wolof	62.6	44.2	54.5	58.9	63.1 (+18.9)
isiXhosa	69.5	24.0	61.7	71.1	70.4 (+46.4)
Yoruba	58.2	36.0	38.1	36.8	41.4 (+5.4)
isiZulu	60.2	43.9	68.9	73.0	74.8 (+30.9)
AVG	60.4	54.5	63.6	64.9	67.1 (+12.7)

- NER: mDeBERTa-v3
- MT: NLLB

prior marker-based approach
cannot do this

Experiment Results

“Translate-test” - CODEC can also translate test data in source language into a high-resource language to run inference on, then project predicted span labels back to the test data.

Lang.	GPT-4 [†]	FT _{En}	Translate-train			Translate-test	
			Awes-align	EasyProject	CODEC (Δ_{FT})	Awes-align	CODEC (Δ_{FT})
Bambara	46.8	37.1	45.0	45.8	45.8 (+8.7)	50.0	55.6 (+18.5)
Ewe	75.5	75.3	78.3	78.5	79.1 (+3.8)	72.5	79.1 (+3.8)
Fon	19.4	49.6	59.3	61.4	65.5 (+15.9)	62.8	61.4 (+11.8)
Hausa	70.7	71.7	72.7	72.2	72.4 (+0.7)	70.0	73.7 (+2.0)
Igbo	51.7	59.3	63.5	65.6	70.9 (+11.6)	77.2	72.8 (+13.5)
Kinyarwanda	59.1	66.4	63.2	71.0	71.2 (+4.8)	64.9	78.0 (+11.6)
Luganda	73.7	75.3	77.7	76.7	77.2 (+1.9)	82.4	82.3 (+7.0)
Luo	55.2	35.8	46.5	50.2	49.6 (+13.8)	52.6	52.9 (+17.1)
Mossi	44.2	45.0	52.2	53.1	55.6 (+10.6)	48.4	50.4 (+5.4)
Chichewa	75.8	79.5	75.1	75.3	76.8 (-2.7)	78.0	76.8 (-2.7)
chiShona	66.8	35.2	69.5	55.9	72.4 (+37.2)	67.0	78.4 (+43.2)
Kiswahili	82.6	87.7	82.4	83.6	83.1 (-4.6)	80.2	81.5 (-6.2)
Setswana	62.0	64.8	73.8	74.0	74.7 (+9.9)	81.4	80.3 (+15.5)
Akan/Twi	52.9	50.1	62.7	65.3	64.6 (+14.5)	72.6	73.5 (+23.4)
Wolof	62.6	44.2	54.5	58.9	63.1 (+18.9)	58.1	67.2 (+23.0)
isiXhosa	69.5	24.0	61.7	71.1	70.4 (+46.4)	52.7	69.2 (+45.2)
Yoruba	58.2	36.0	38.1	36.8	41.4 (+5.4)	49.1	58.0 (+22.0)
isiZulu	60.2	43.9	68.9	73.0	74.8 (+30.9)	64.1	76.9 (+33.0)
AVG	60.4	54.5	63.6	64.9	67.1 (+12.7)	65.8	70.4 (+16.0)

Error Analysis

Underline marks the projection errors.



	English Data	EasyProject	Awesome-align	Codec
chiShona	India _{LOC} and Pakistan _{LOC} have fought ... region of Kashmir _{LOC} ...	India _{LOC} <u>ne</u> Pakistan _{LOC} ... ye Kashmir _{LOC} chibviro ...	India _{LOC} <u>ne</u> Pakistan ... zvinetso <u>ye</u> Kashmir _{LOC} ...	India _{LOC} nePakistan _{LOC} ... zvinetso yeKashmir _{LOC} ...
isiZulu	State media quoted China _{LOC} 's top negotiator with Taipei _{LOC} , Tang Shubei _{PER} , ... from Taiwan _{LOC} ...	Imithombo ... <u>we</u> China _{LOC} <u>ne</u> Taipei _{LOC} , uTang Shubei _{PER} , ... elivela eTaiwan _{LOC} ...	Imithombo _{LOC} ... <u>wase</u> China _{LOC} <u>ne</u> Taipei _{LOC} , uTang Shubei _{PER} , ... elivela eTaiwan _{LOC} ...	Imithombo ... waseChina _{LOC} <u>ne</u> Taipei _{LOC} , uTang Shubei _{PER} , ... elivela eTaiwan _{LOC} ...

only marks sub-words
as an entity

Augmented data in low-resource languages

having difficulty
to project multiple spans

Today's talk — three social aspects of LLMs

1 - Cultural Biases



(Naous et al., ACL 2024)

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

2 - World Languages

CODEC



(Le et al., ICLR 2024)

Design decoding
algorithms to improve
performance on
non-English languages.

3 - User Privacy



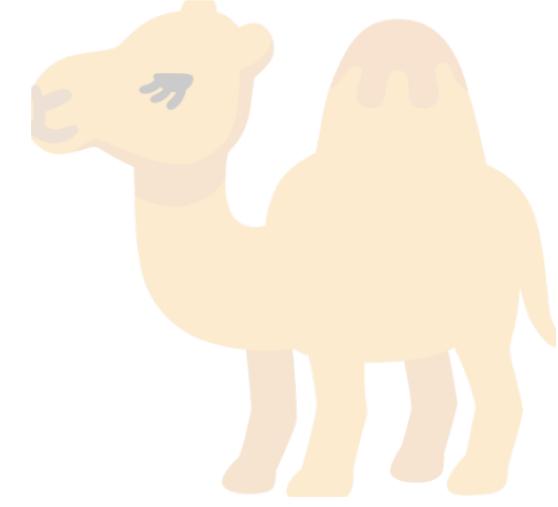
(Yao et al., ACL 2024)

Democratize the
privacy protection via
human-centered AI to
empower end users.

Today's talk — three social aspects of LLMs

1 - Cultural Biases

CAMEL



(Naous et al., ACL 2024)

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

2 - World Languages

CODEC



(Le et al., ICLR 2024)

Design decoding
algorithms to improve
performance on
non-English languages.

3 - User Privacy

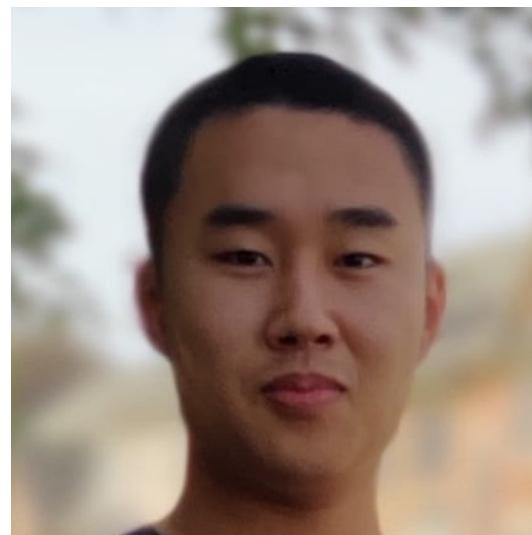
PrivacyMirror



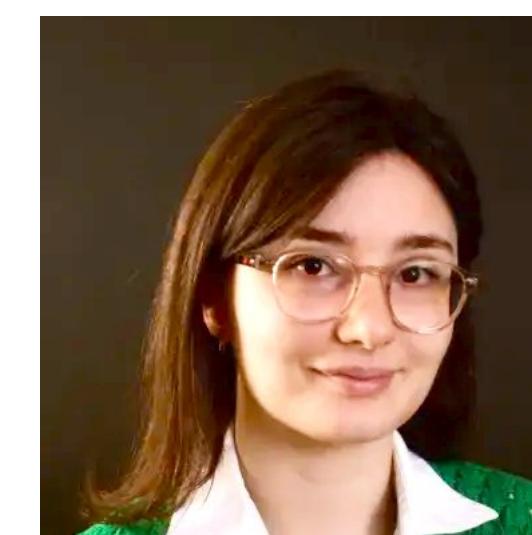
(Yao et al., ACL 2024)

Democratize the
privacy protection via
human-centered AI to
empower end users.

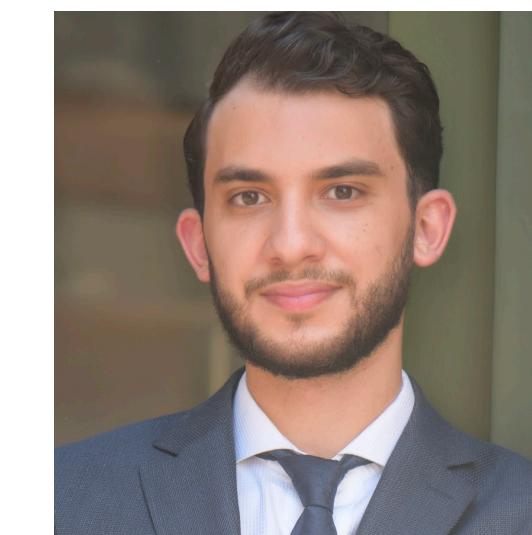
Reducing Privacy Risks in Online Self-Disclosures (PrivacyMirror



Yao Dou



Isadora Krsek



Tarek Naous



Anubha Kabra



Sauvik Das



Alan Ritter



Wei Xu

People talk about themselves online

Or, send information about themselves or others to the LLMs 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 ...

People talk about themselves online

Or, send information about themselves or others to the LLMs 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 ...

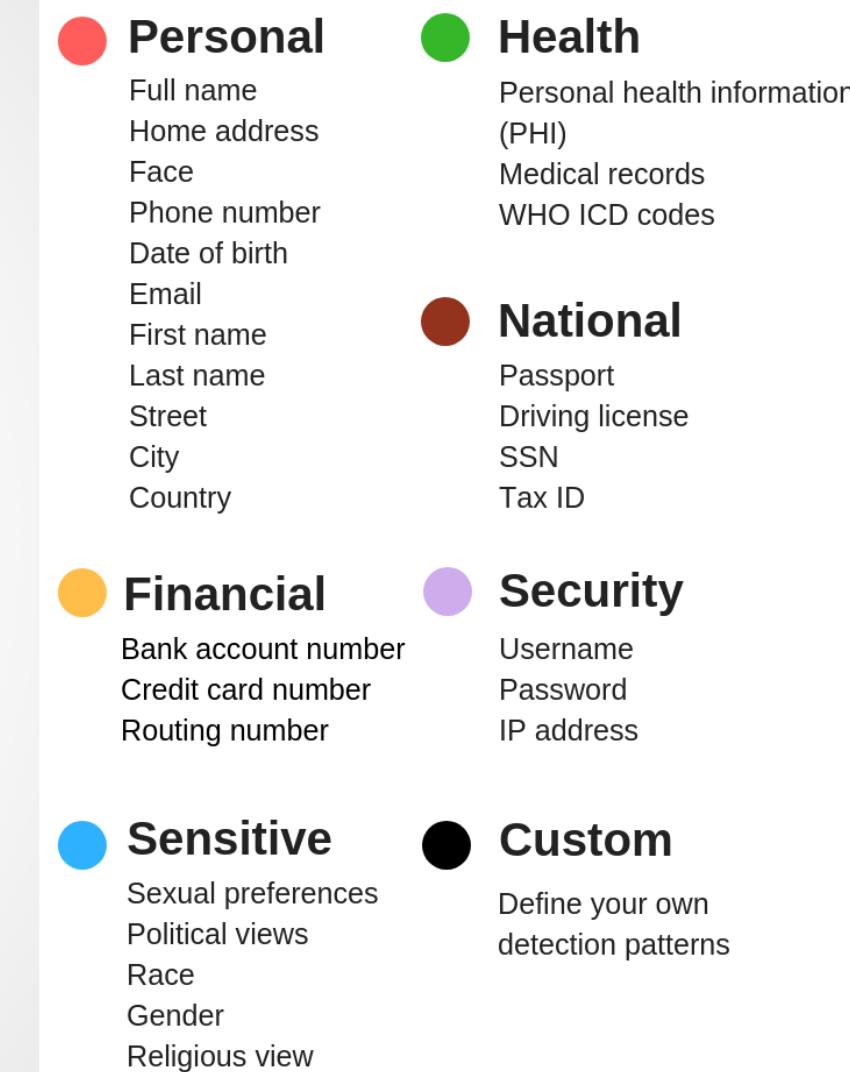
Disclosures:

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



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

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



PrivacyMirror — 19 Self-disclosure Categories

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

Demographic Attributes

Age Wife/GF

Age&Gender Husband/BF

Race/Nationality Sexual Orientation

Gender Relationship Status

Location Pet

Appearance 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.

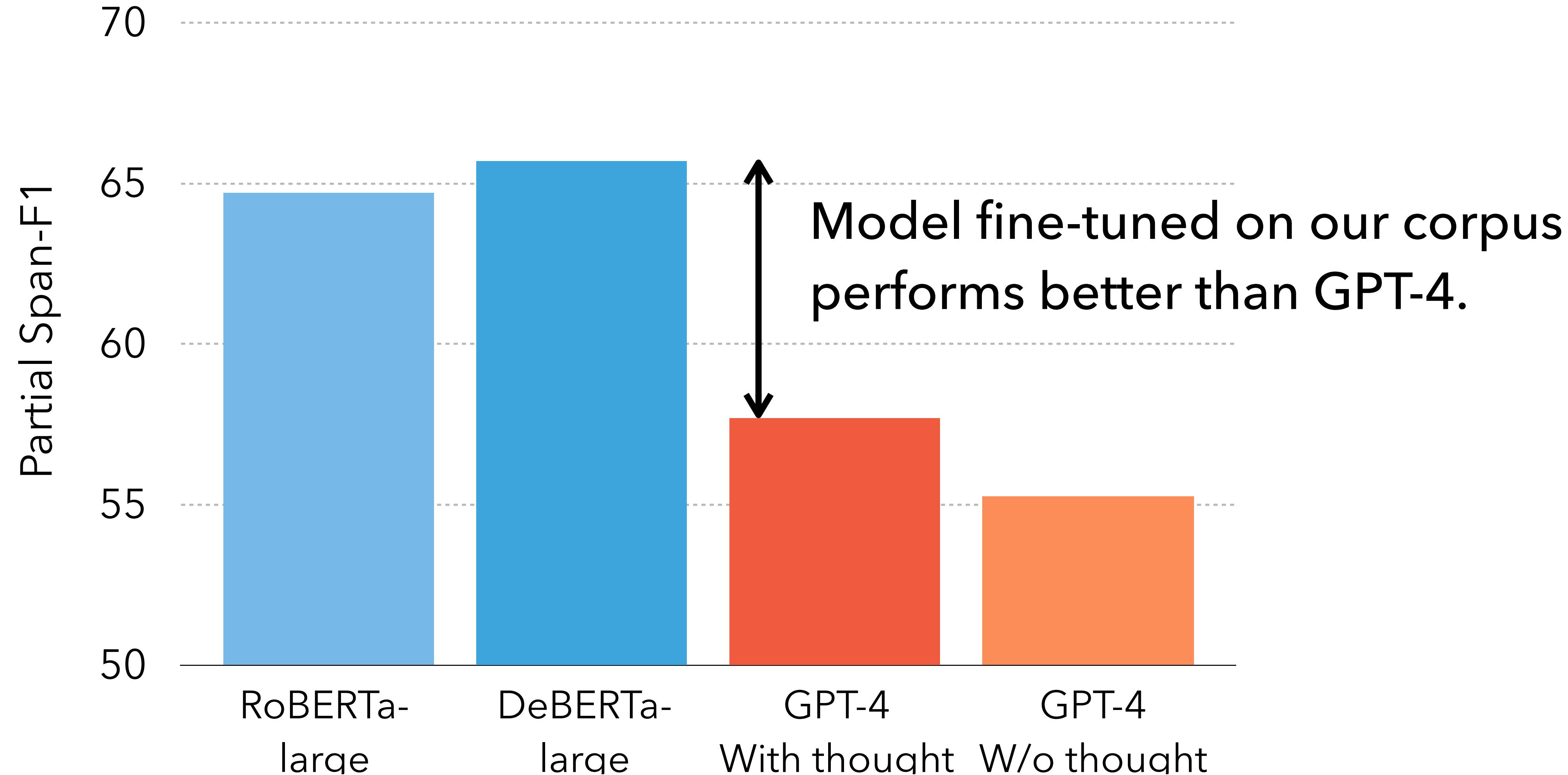
My little brother (9M) is my pride and joy

My husband and I vote for different parties



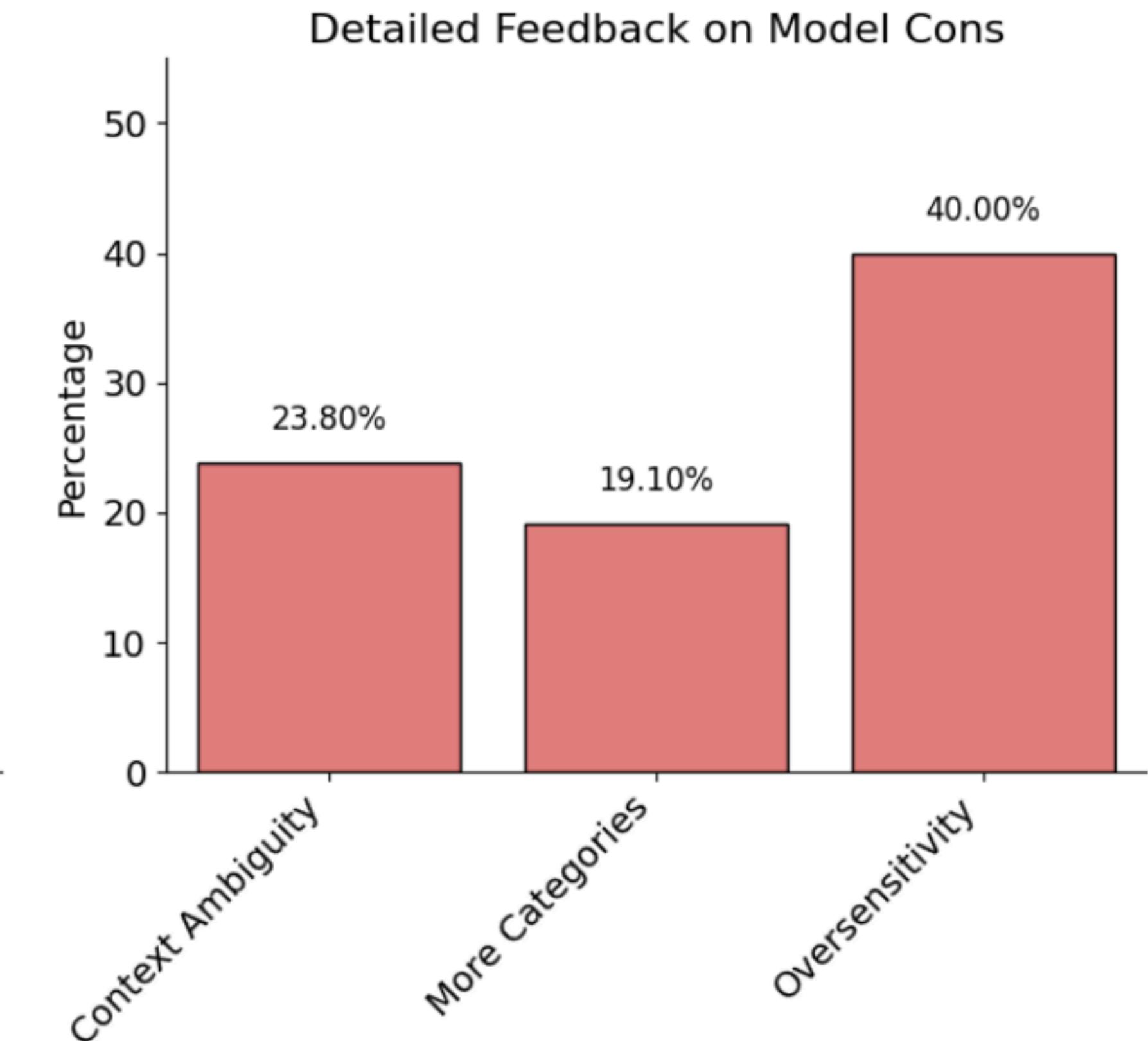
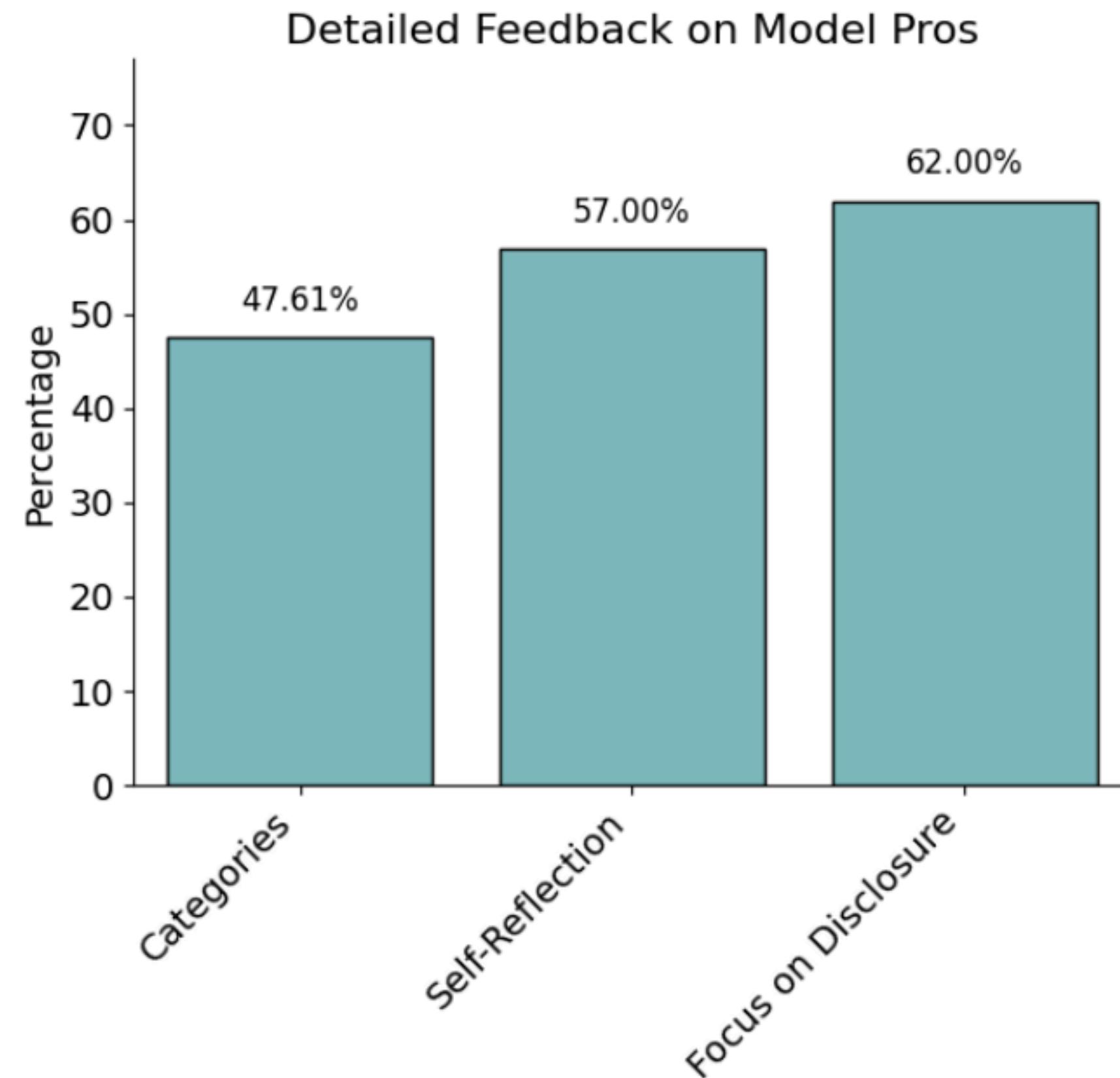
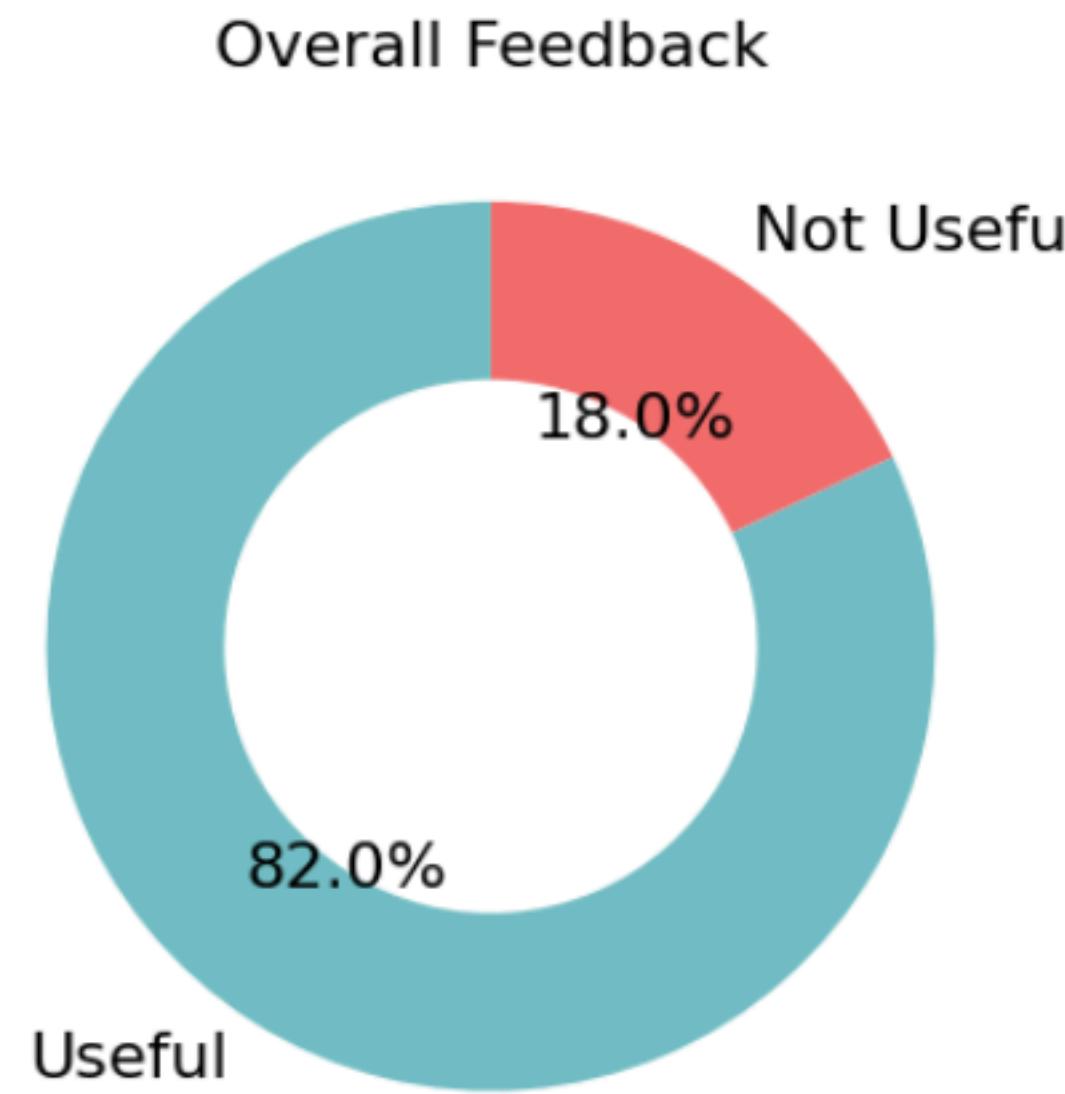
PrivacyMirror — Self-disclosure Detection

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



Do real users like our detection model?

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.





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 about 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] X Utility

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

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

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

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

✗ Utility

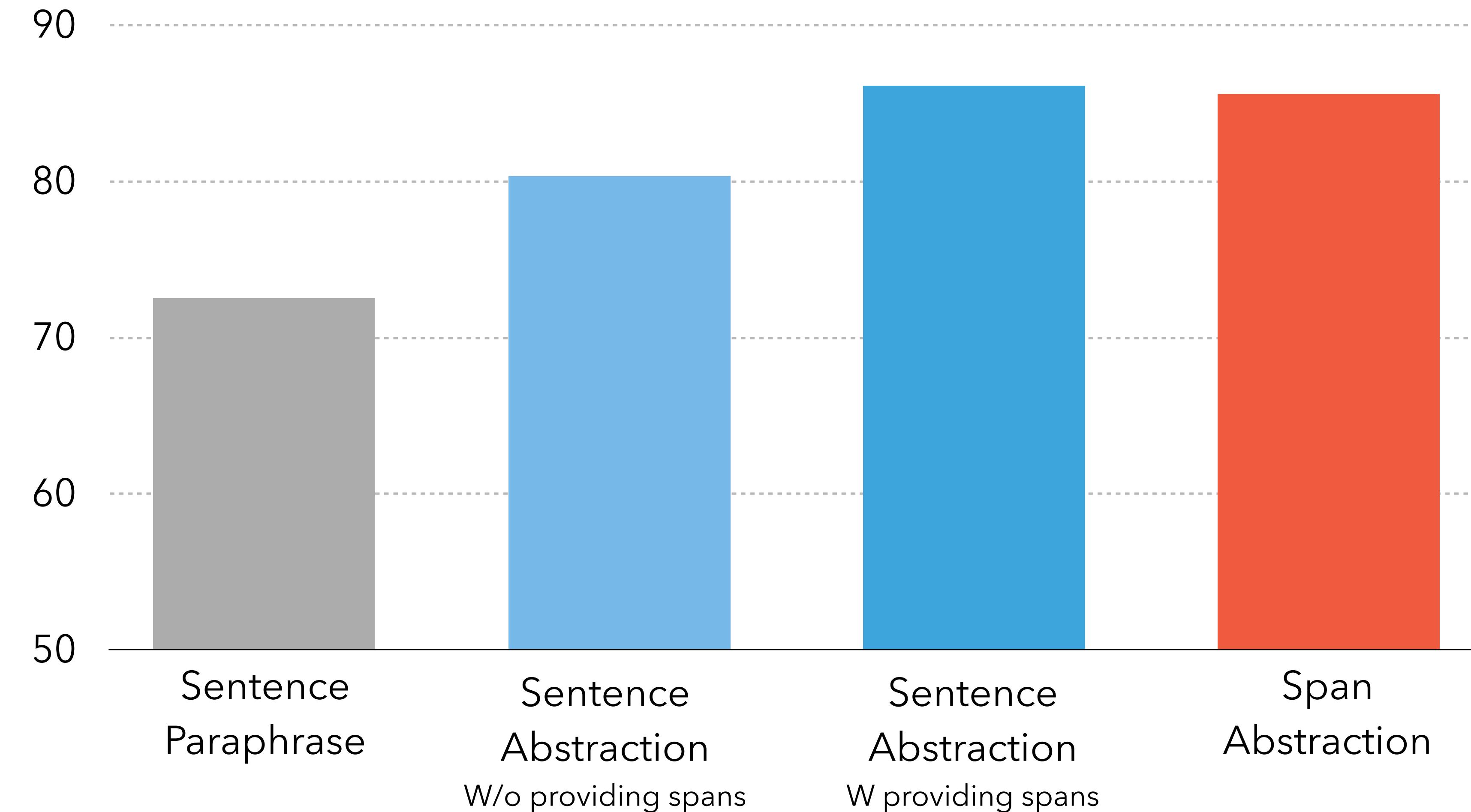
✗ Privacy

✗ Writing Style



PrivacyMirror — Self-disclosure Abstraction

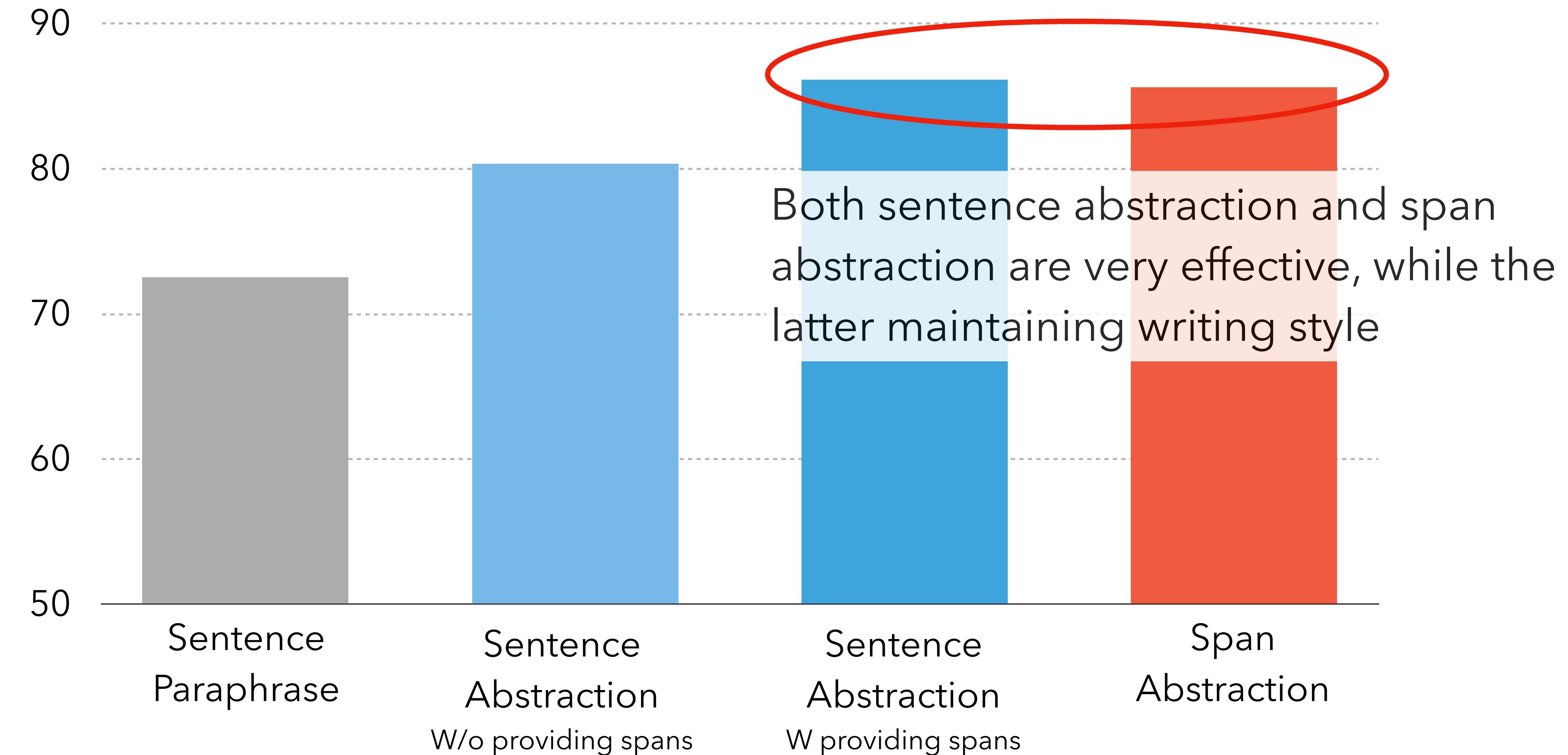
Human evaluation on effectiveness (consider both utility preservation & privacy increase) w/ GPT-4





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.
- Fine-tuning LLMs to detect self-disclosures is feasible but has room for improvements;
- Fine-tuning LLMs to abstract disclosures works pretty well.

Paper on arXiv

The screenshot shows the arXiv preprint page for the paper. The title is "Reducing Privacy Risks in Online Self-Disclosure with Language Models". The authors listed are Yao Dou[†], Isadora Krsek^e, Tarek Naous^π, Anubha Kabra^e, Sauvik Das^e, Alan Ritter^π, Wei Xu^π. Affiliations: [†]Georgia Institute of Technology, ^eCarnegie Mellon University. The email address douy@gatech.edu is shown. Below the title is the abstract section, which includes a diagram illustrating the process of abstracting self-disclosures. The diagram shows a user input like "I'm 16F I think I want to be a bi M" being processed by a language model to produce an abstracted version like "I am exploring my sexual identity". Other examples show the model abstracting details about living in New Mexico, family members, and past experiences.

[cs.CL] 20 Feb 2024

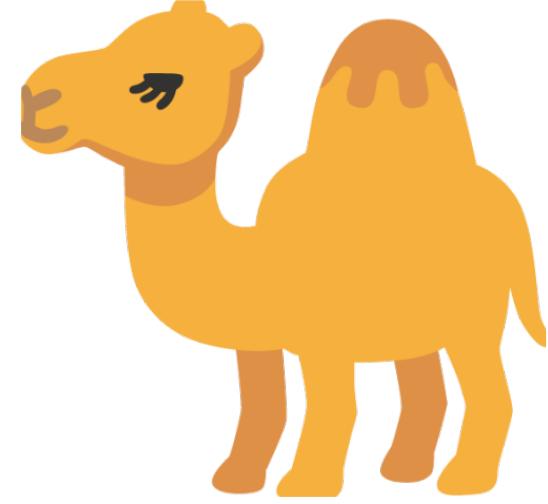
Model on Huggingface

The screenshot shows the Huggingface Model Card for the Llama-2-7B-lora-instruction-ft-abstraction-three-span model. The card includes the author's profile (douy), the model name, a brief description ("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.'"), and links to the paper (arXiv:2311.09538, arXiv:2106.09685), files, and community discussions. A sidebar on the right provides navigation links for Models and Datasets.

Today's talk — three social aspects of LLMs

1 - Cultural Biases

CAMEL



(Naous et al., ACL 2024)

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

2 - World Languages

CODEC



(Le et al., ICLR 2024)

Design decoding algorithms to improve performance on non-English languages.

3 - User Privacy

PrivacyMirror



(Yao et al., ACL 2024)

Democratize the privacy protection via human-centered AI to empower end users.

Conclusions



We need not only multilingual LLMs, but also multicultural LLMs.



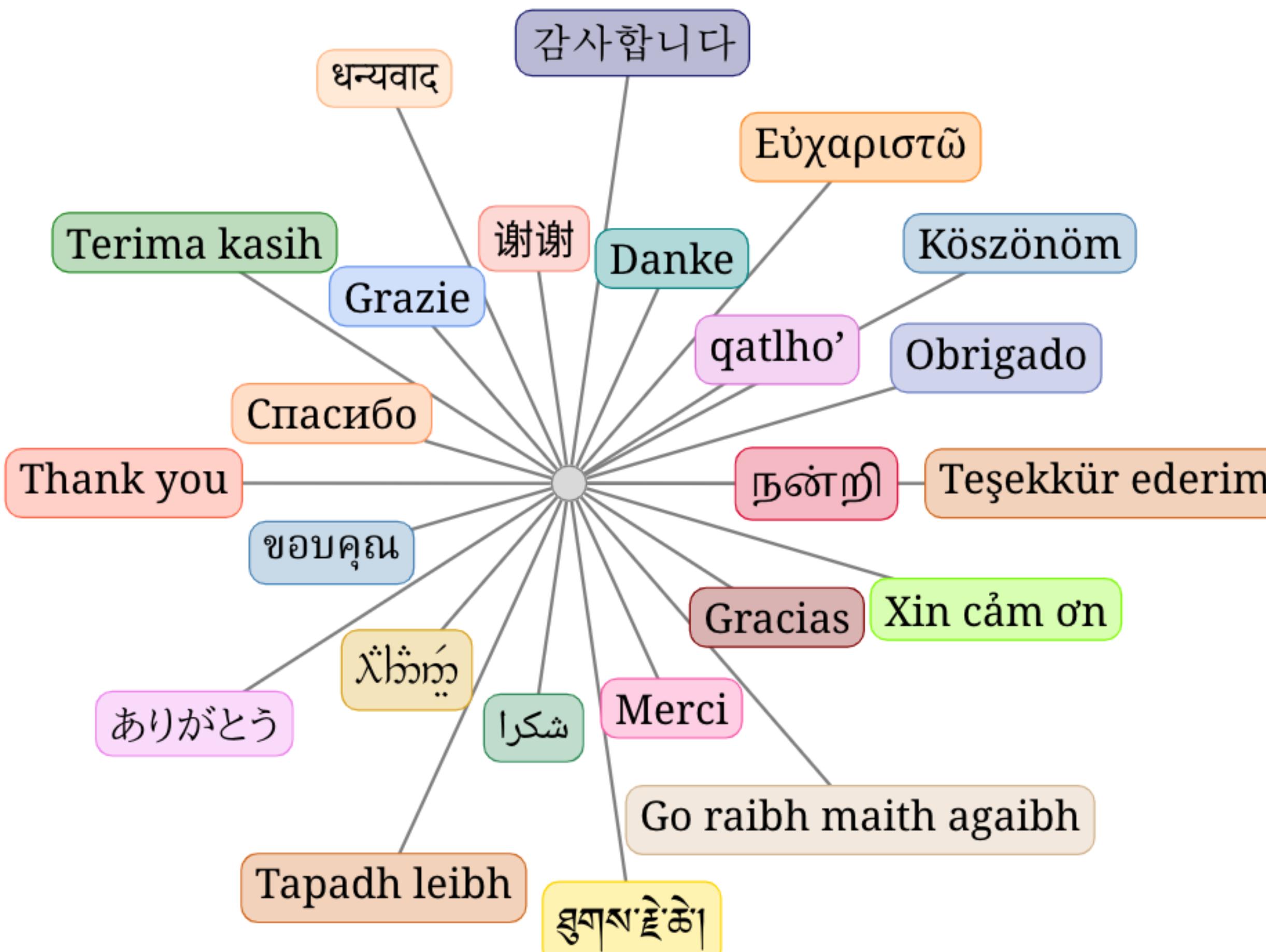
How we sample, how we handle pre-training data is very important for deployment of LLMs worldwide. Decoding algorithms can also make a big difference.



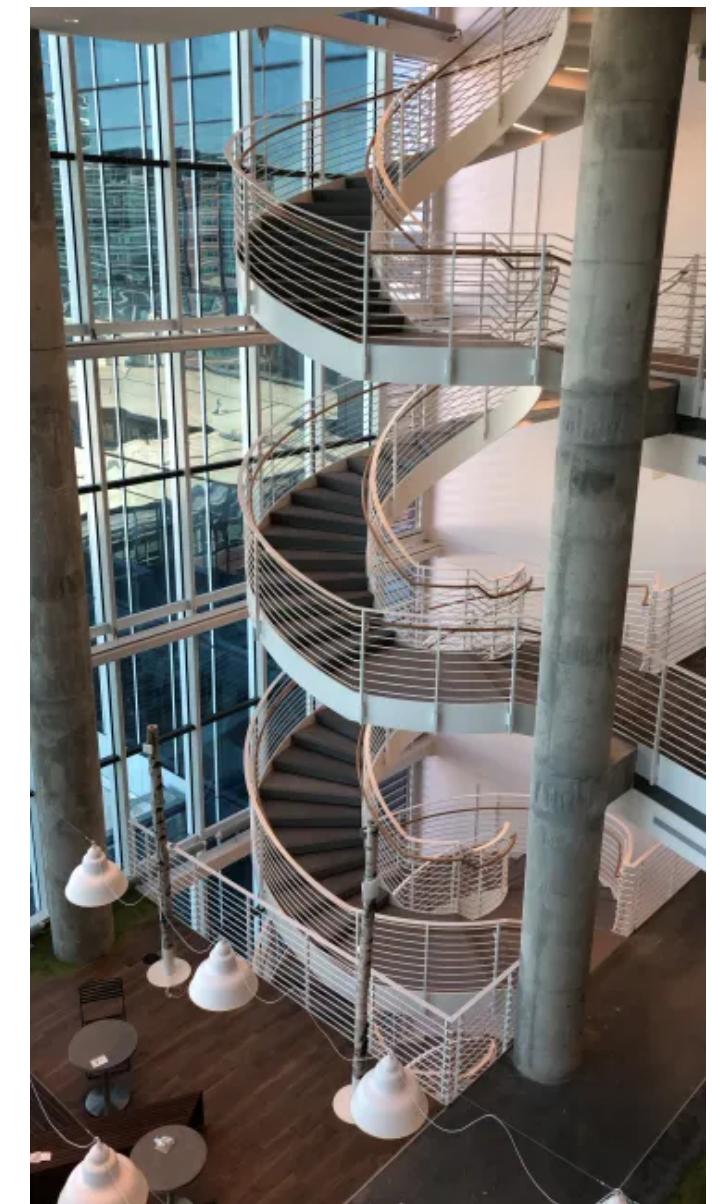
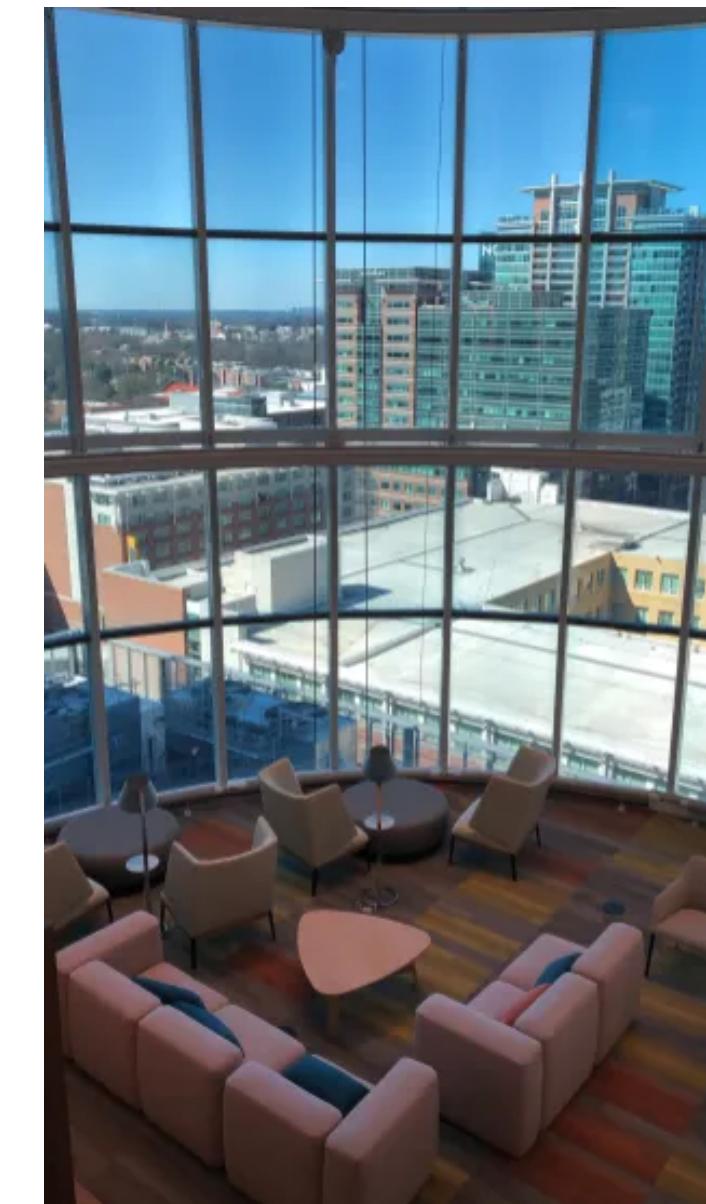
We will want to democratize the privacy protection, empowering individual end users to protect their own data.

Thank you!

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



(image credit: Overleaf)



(image credit: Georgia Tech)



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

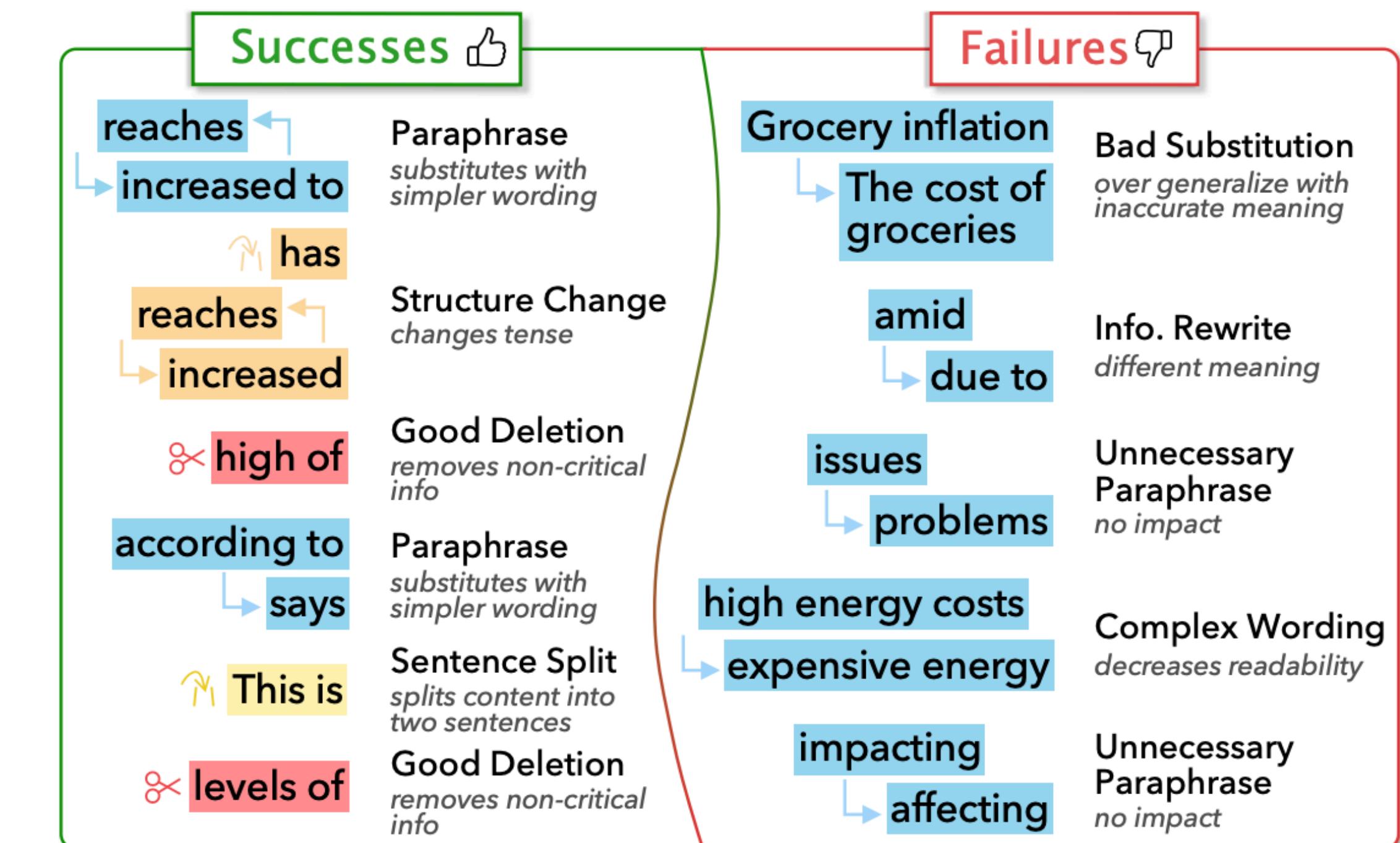
Here is another example of text simplification. GPT-4 rewrites complex text into simpler language.

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:

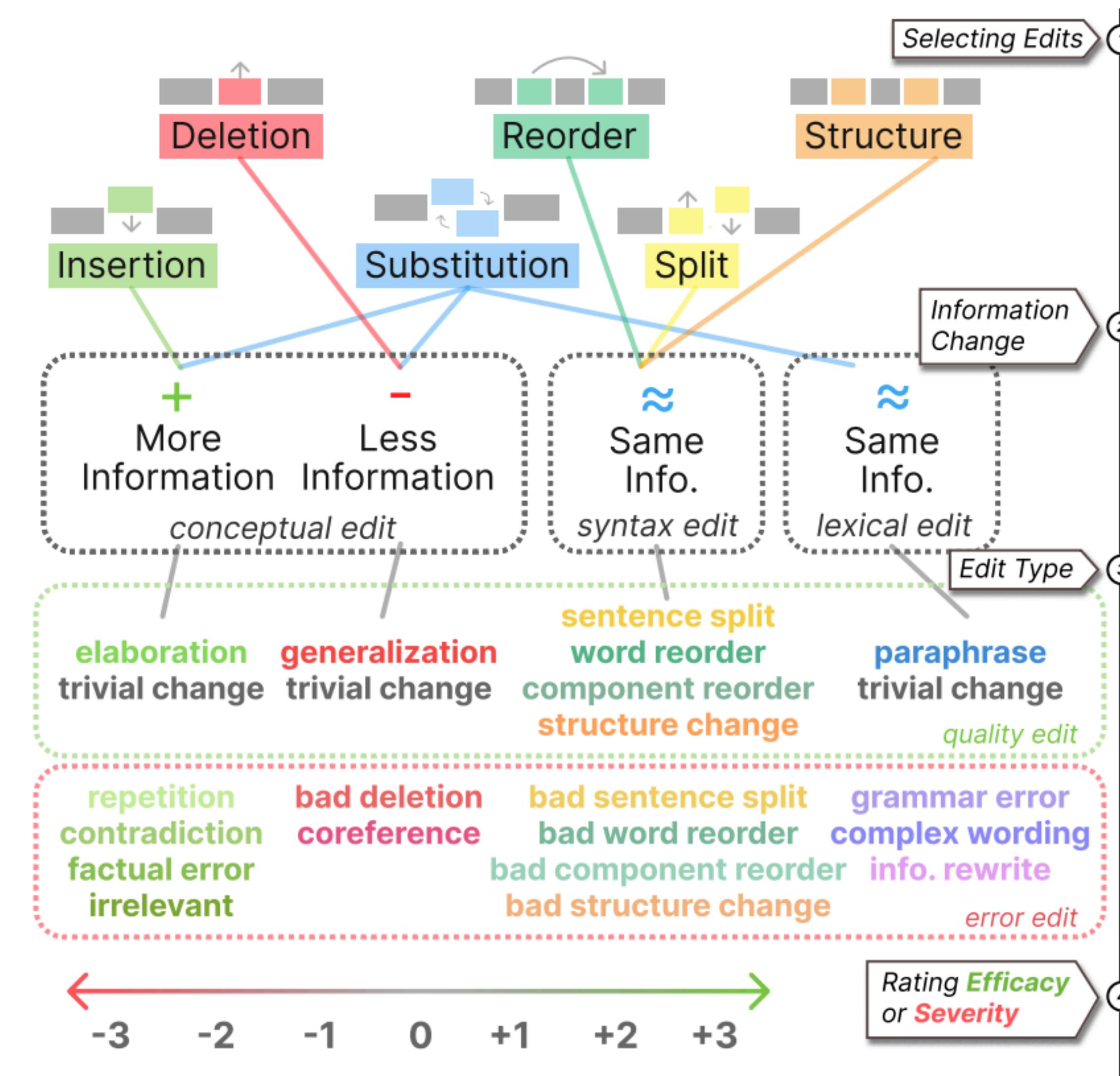
The cost of groceries in the United Kingdom has increased to a record 17.1%, says market research group Kantar Worldpanel. || This is due to high inflation, supply chain problems, and expensive energy affecting the economy.



Errors in LLM-generated texts can be difficult to capture

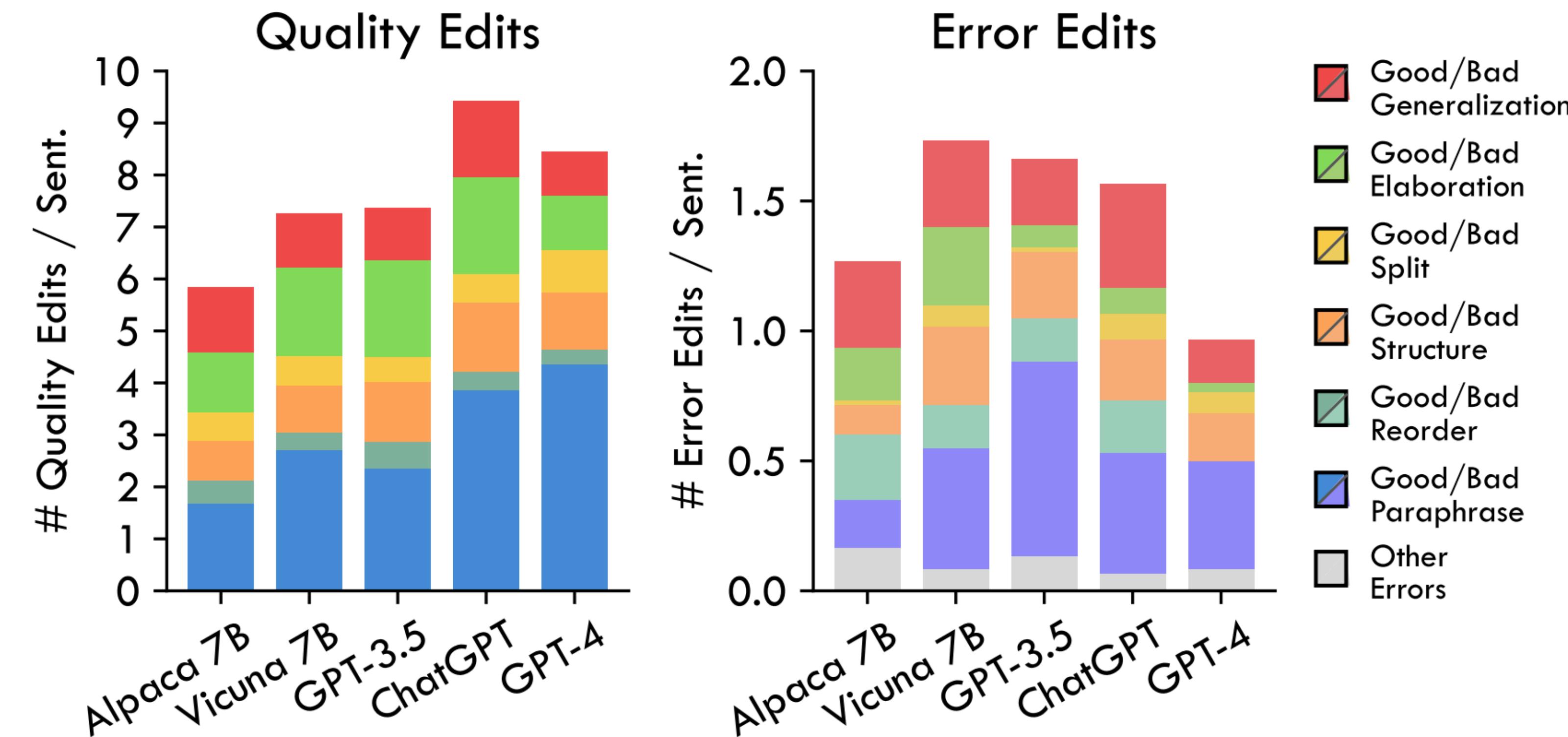
Thresh — typology for edit-level evaluation

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.

Evaluating Robustness of Large Language Models with Neologisms (NeoBench)



Jonathan Zheng



Alan Ritter



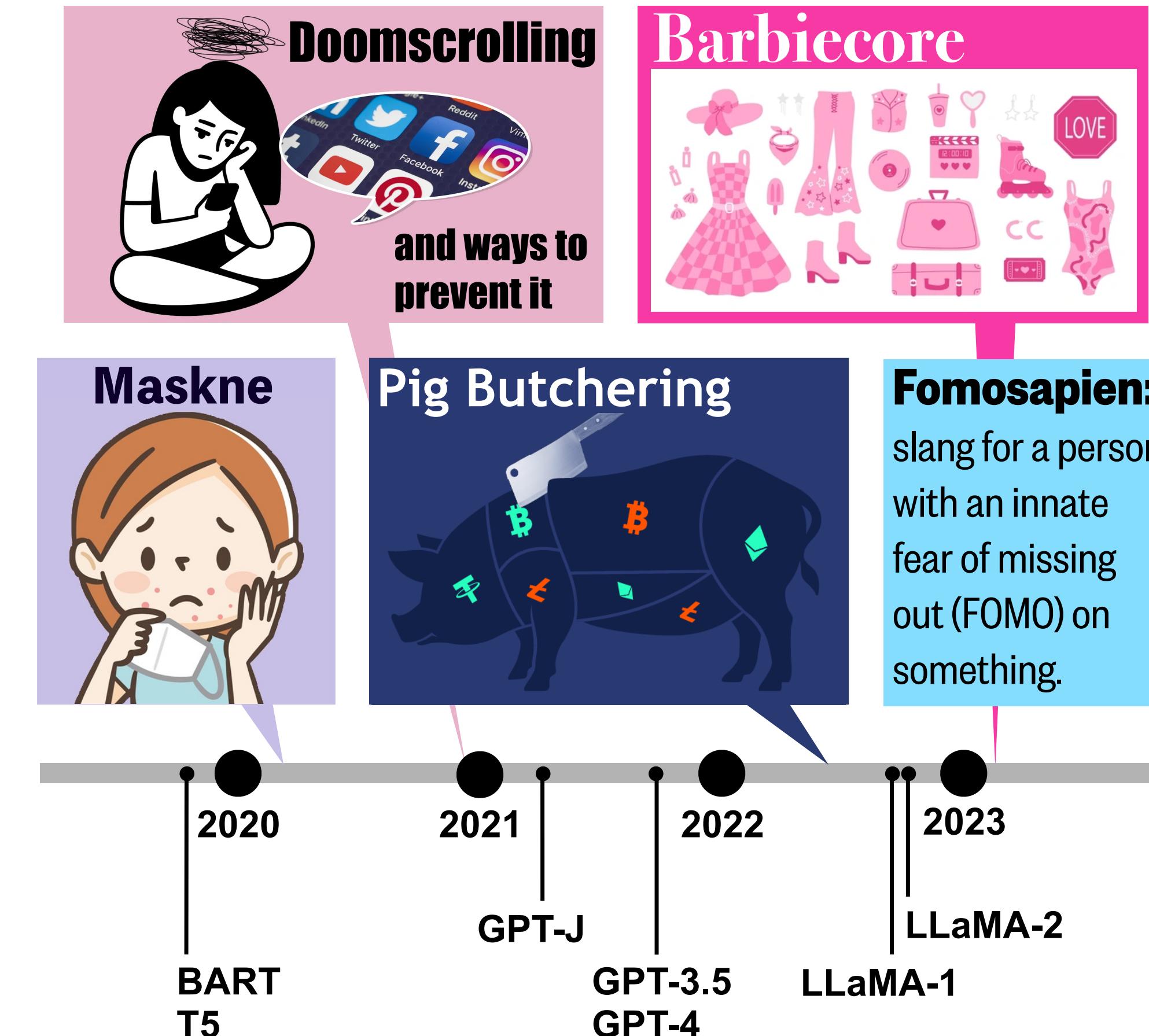
Wei Xu

A better technical solution for
marker-based label projection



NeoBench — evolving human languages

Data contamination, long-tail low-frequency words, tokenization, ...



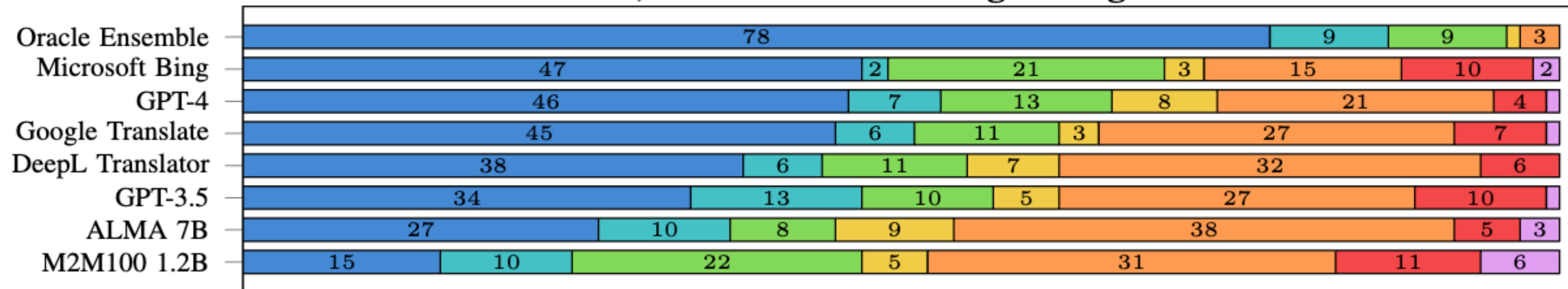
We used 3 different methods to obtain 2,505 single- and multi-word neologisms.



NeoBench — human evaluation on translation

Models struggle to translate sentences that contains neologism (vs. non-neologism) word.

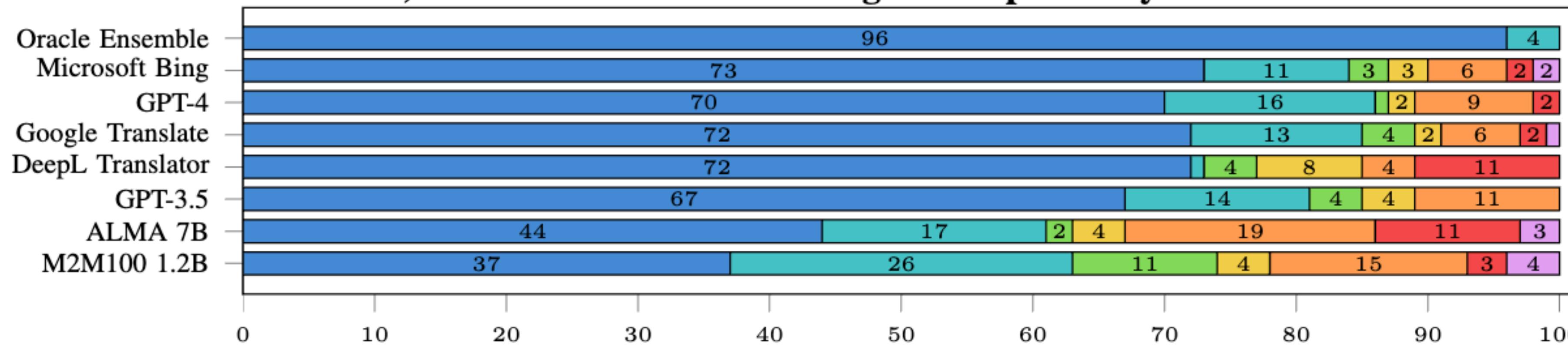
a) Sentences Containing Neologisms



Example:

Starting to think **doomscrolling** through the fall of civilization is having a negative effect on my mental health.

b) Same Sentences with Neologisms Replaced by Common Words



Example:

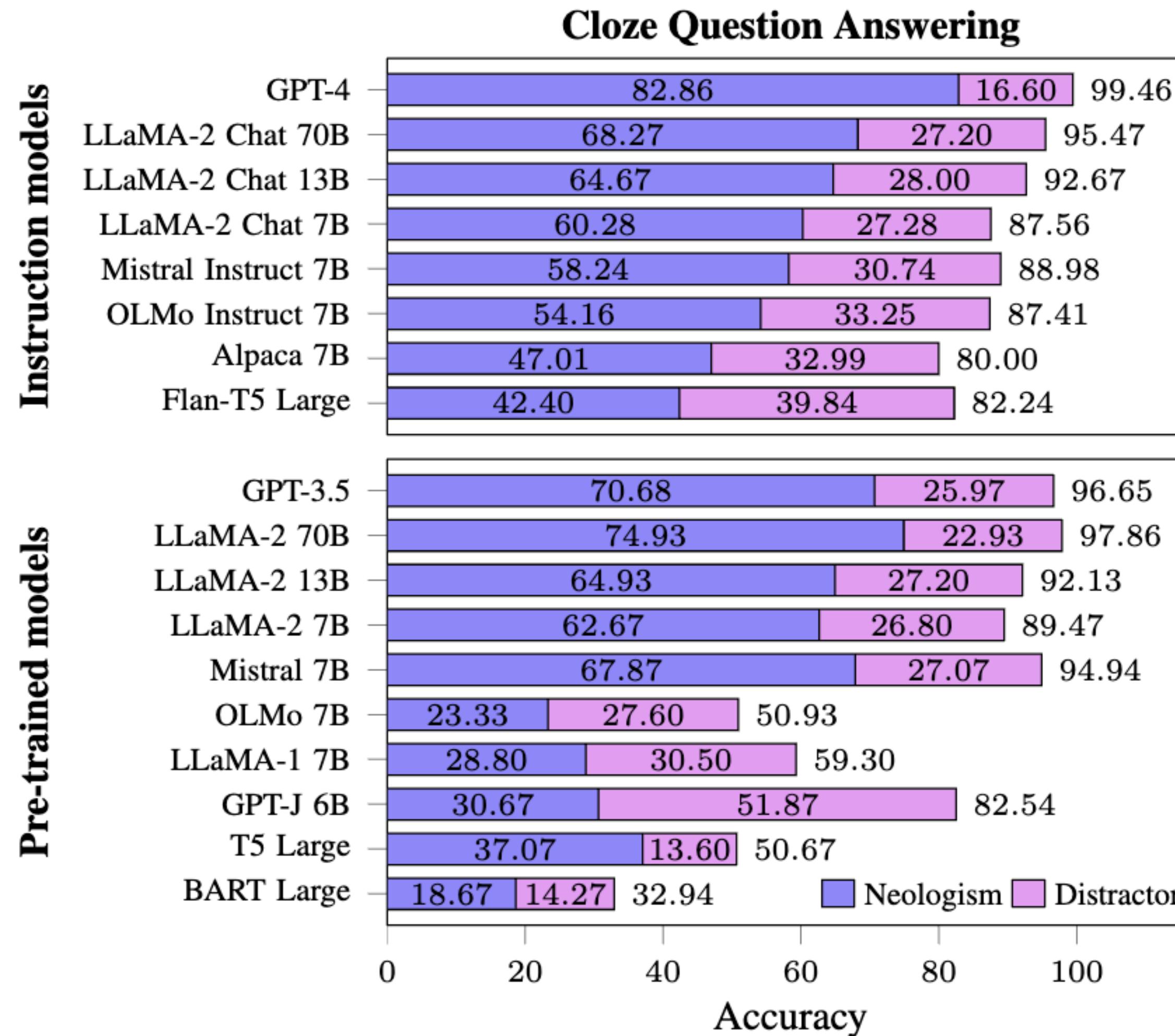
Starting to think **smoking** through the fall of civilization is having a negative effect on my mental health.

Legend: Good (blue), Unnatural (teal), Literal (green), Partial (yellow), Mistranslation (orange), Copy (red), Incomprehensible (purple)



NeoBench — perplexity, Cloze QA, definition

Newer, larger LLMs work better; but, perplexity becomes worse after instruction tuning.



Neologism: *doomscrolling*

The silver lining of this website no longer functioning as an even vaguely reliable information source is that ___ has basically been completely undermined. It wouldn't even work now since everything is too geared to outrage clickbait and actual reporting has disappeared, so there is no point staying on the app.

- | | |
|-------------------------|-----------------------------------|
| a) misinformation | b) surfing |
| c) doomscrolling | d) lying |
| e) gaming | f) anti-productivity (distractor) |
-

Table 2: Example passage in NEO-BENCH for multiple-choice Cloze Question Answering with correct neologism answers and partially correct distractor answers.



NeoBench — perplexity, Cloze QA, definition

Newer, larger LLMs work better; but, perplexity becomes worse after instruction tuning.

