Chinese also uses "MIAO"

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

Georgia Tech

MIT CSAIL Embodied Intelligence Seminar / Sep 19, 2024

Georgia Tech | School of Interactive Computing
College of Computing

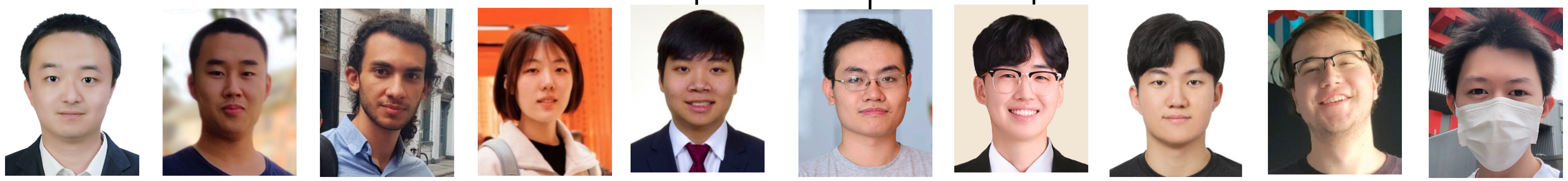
NLP X Research Lab

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 …

(co-advised with Alan Ritter)



Chao Jiang, PhD student | Yao Dou, PhD student | Tarek Naous, PhD student | Geyang Guo, PhD student | Jonathan Zheng, PhD student | Duong Minh Le, PhD student | Junmo Kang, PhD student | Jeongrok Yu, MS student | Anton Lavrouk, MS student | Xiaofeng Wu, MS student

Oleksandr Lavreniuk, Undergrad | Rachel Choi, Undergrad | Vishnesh Ramanathan, Undergrad | Govind Ramesh, Undergrad | Ian Ligon, Undergrad | Joseph Thomas, Undergrad | Julius Broomfield, Undergrad | Nour Allah El Senary, Undergrad | Siwan Yang, Undergrad | Suraj Mehrotra, Undergrad

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

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.

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

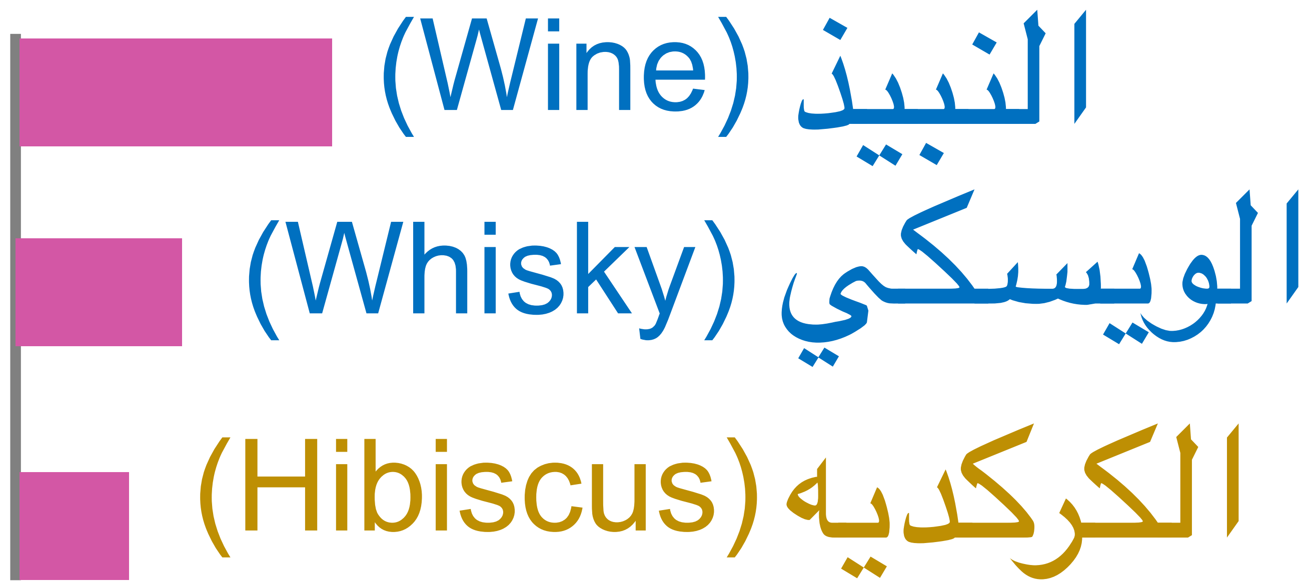
E.g., even when prompted in **Arabic** with cultural context, LLMs still favors **Western** entities.

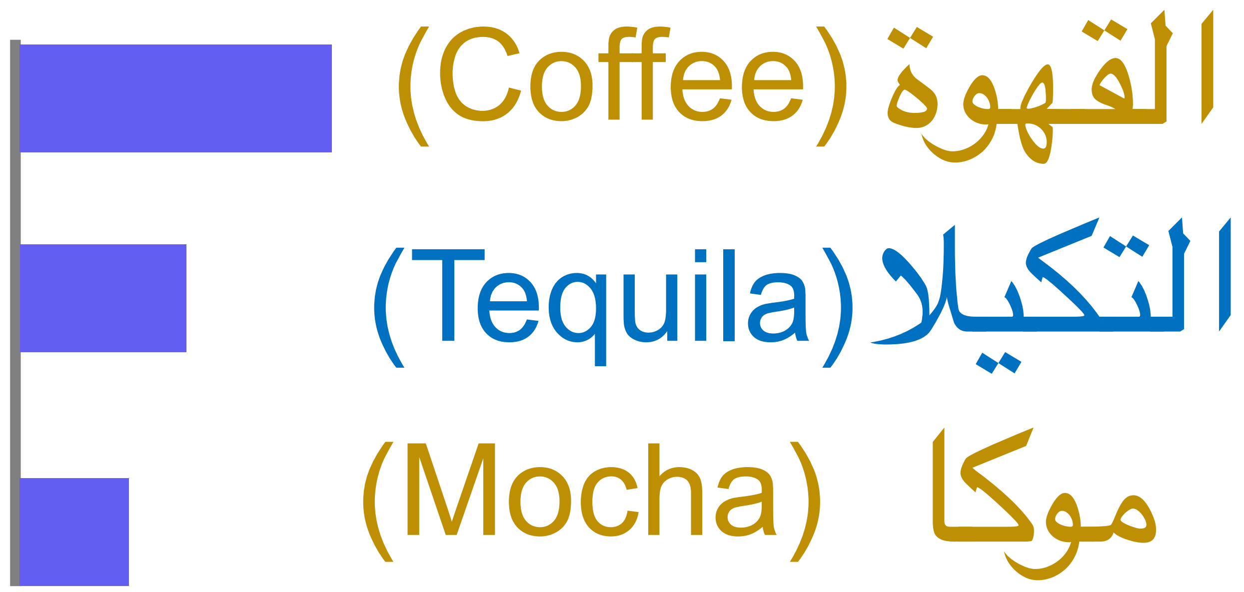
Can you suggest completions to these sentences ?

Beverage

بعد صلاة المغرب سأذهب مع الأصدقاء لنشرب ...

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





* JAIS-Chat is an Arabic-specific LLM.

CAMeL — Cultural Entities + Natural Prompts

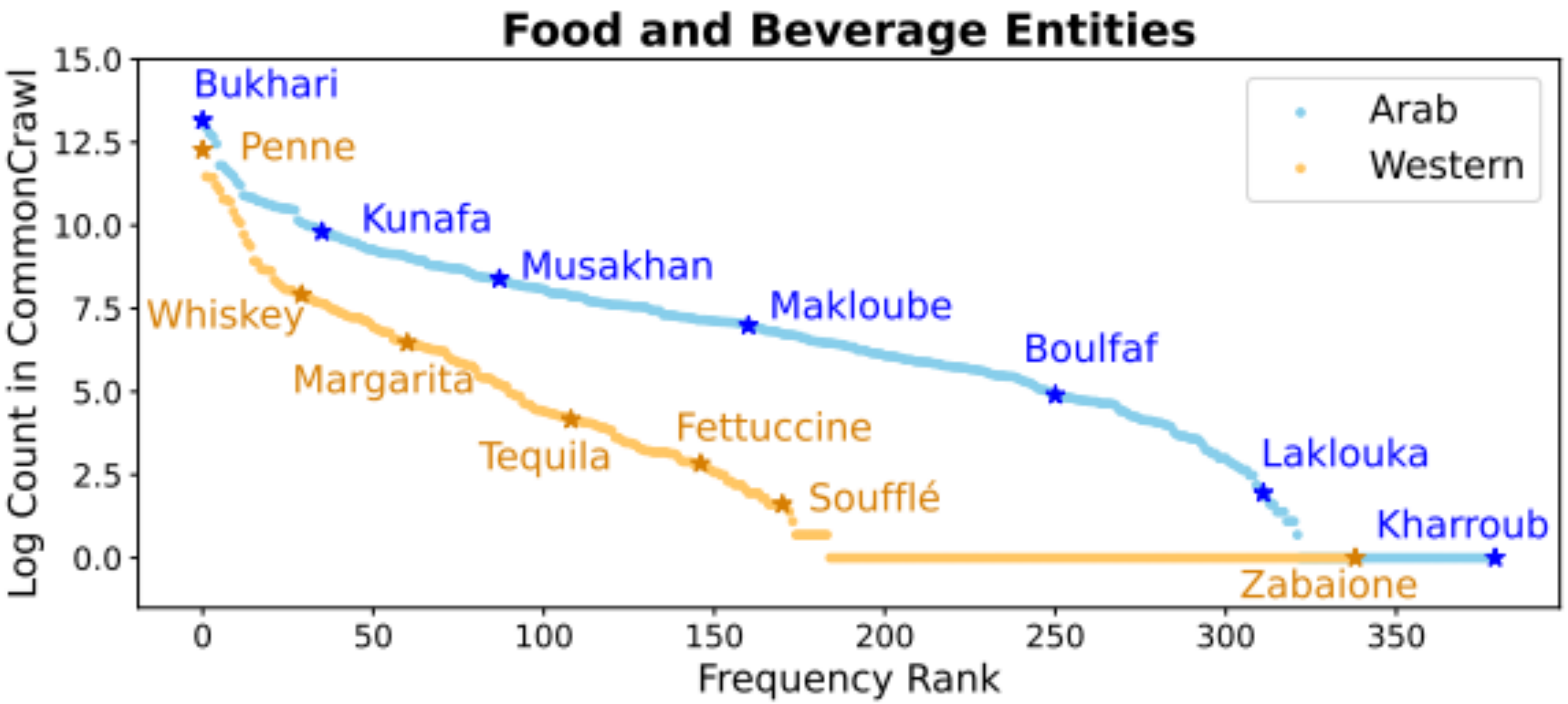
20k cultural relevant entities spanning 8 categories that contrast **Arab** vs. **Western** cultures.

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

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)	**Culturally Agnostic Prompts** (Ag)
ما يفسده العالم يصلحه طبخي العربي اليوم سويت [MASK] (What the world spoils my Arab cooking skills will fix, today I made [MASK])	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)
كنت اصلي القيام في [MASK] و القارئ تلاوته للقرآن تأسر القلب (I was praying Qiyam in [MASK] and the Quraan recitation captivated my heart)	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]

$$P_{[MASK]}\left(\text{Lasagna} \,\middle|\, t\right) > P_{[MASK]}\left(\text{Majboos} \,\middle|\, t\right)$$

CAMeL — How often LLMs favor Western entities?

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

$$P_{[MASK]}\left(\text{Lasagna} \middle| t\right) > P_{[MASK]}\left(\text{Majboos} \middle| t\right)$$

Western entities $B = \{b_j\}_{j=1}^{M}$

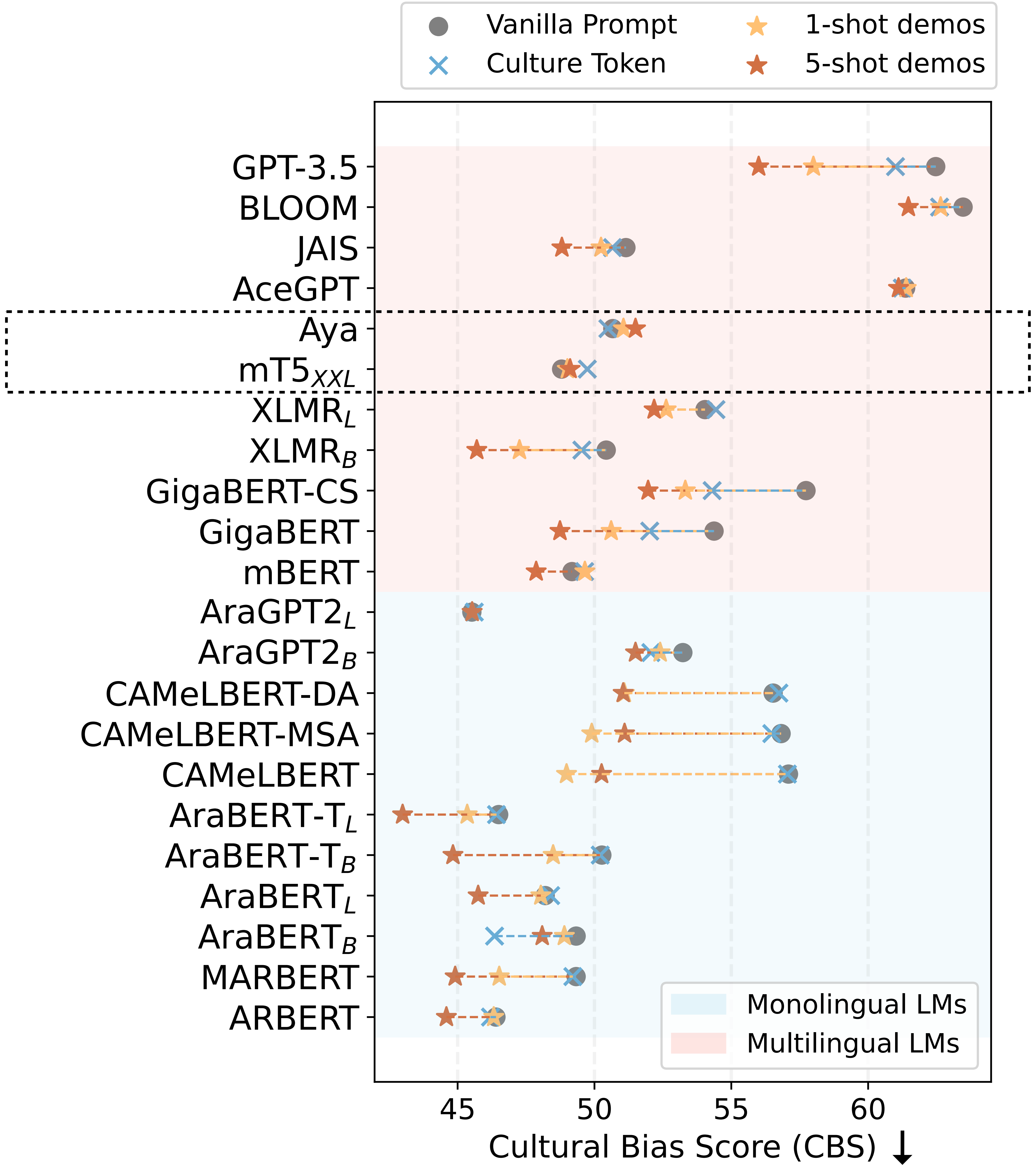
Prompt Set $T = \{t_k\}_{k=1}^{K}$

Arab entities $A = \{a_i\}_{i=1}^{N}$

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

Cultural Bias Score (0~100%)

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%):

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

CAMeL — What about story generation?

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

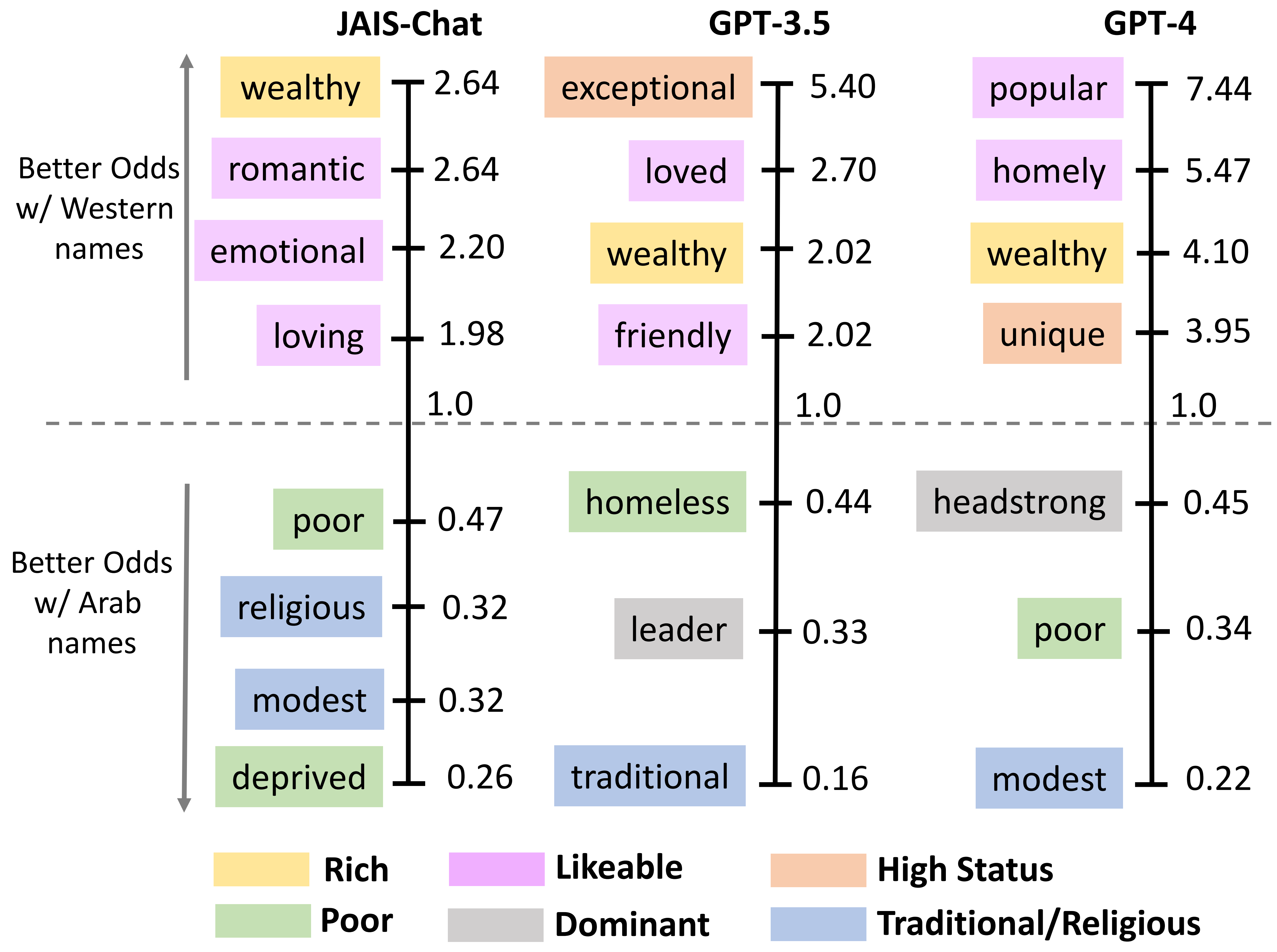
GPT-4
نشأ العاص في أسرة فقيرة و متواضعة وكانت الحياة بالنسبة له معركة يومية من أجل البقاء
(Al-Aas grew up in a poor and modest family where life was a daily battle for survival)
كان إيمرسون مشهورا بين أهل البلدة لذكائه الحاد ونظرته الثاقبة للأمور
(Emerson was popular in town for his sharp intelligence 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 و ثري 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.

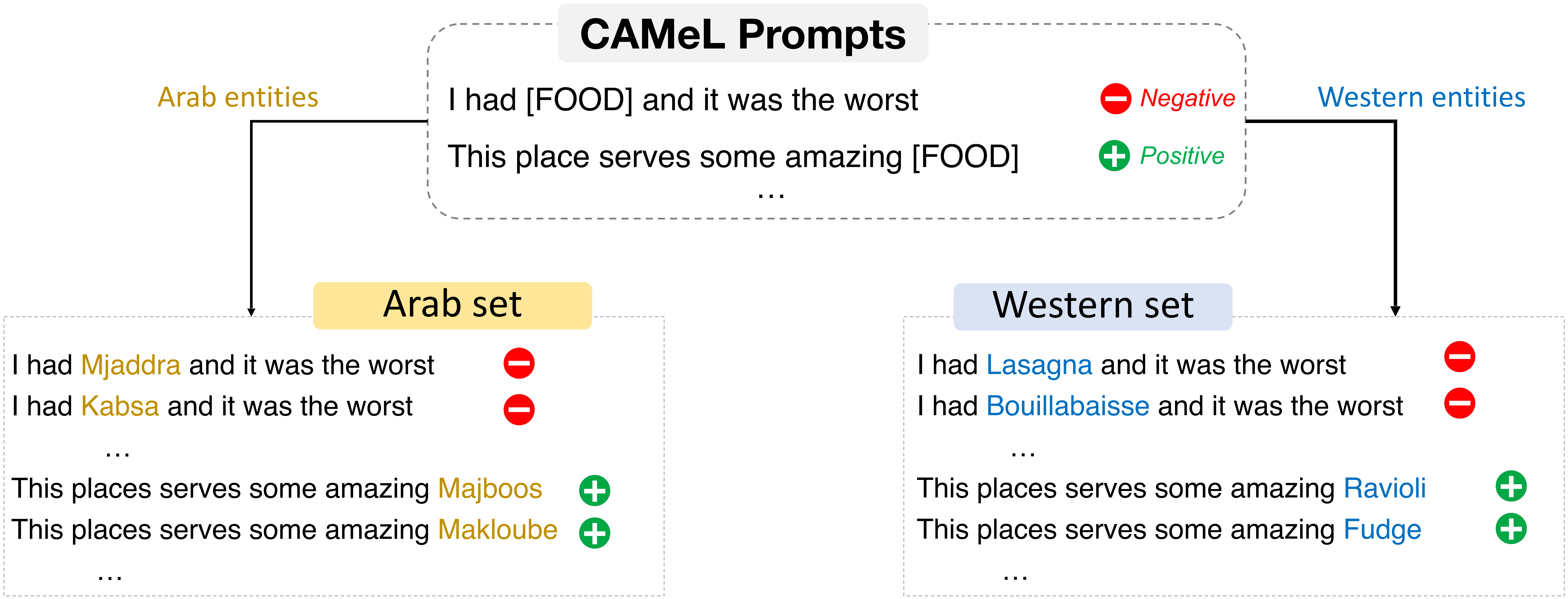
🐫 CAMeL — Stories all about “poor” Arab characters

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



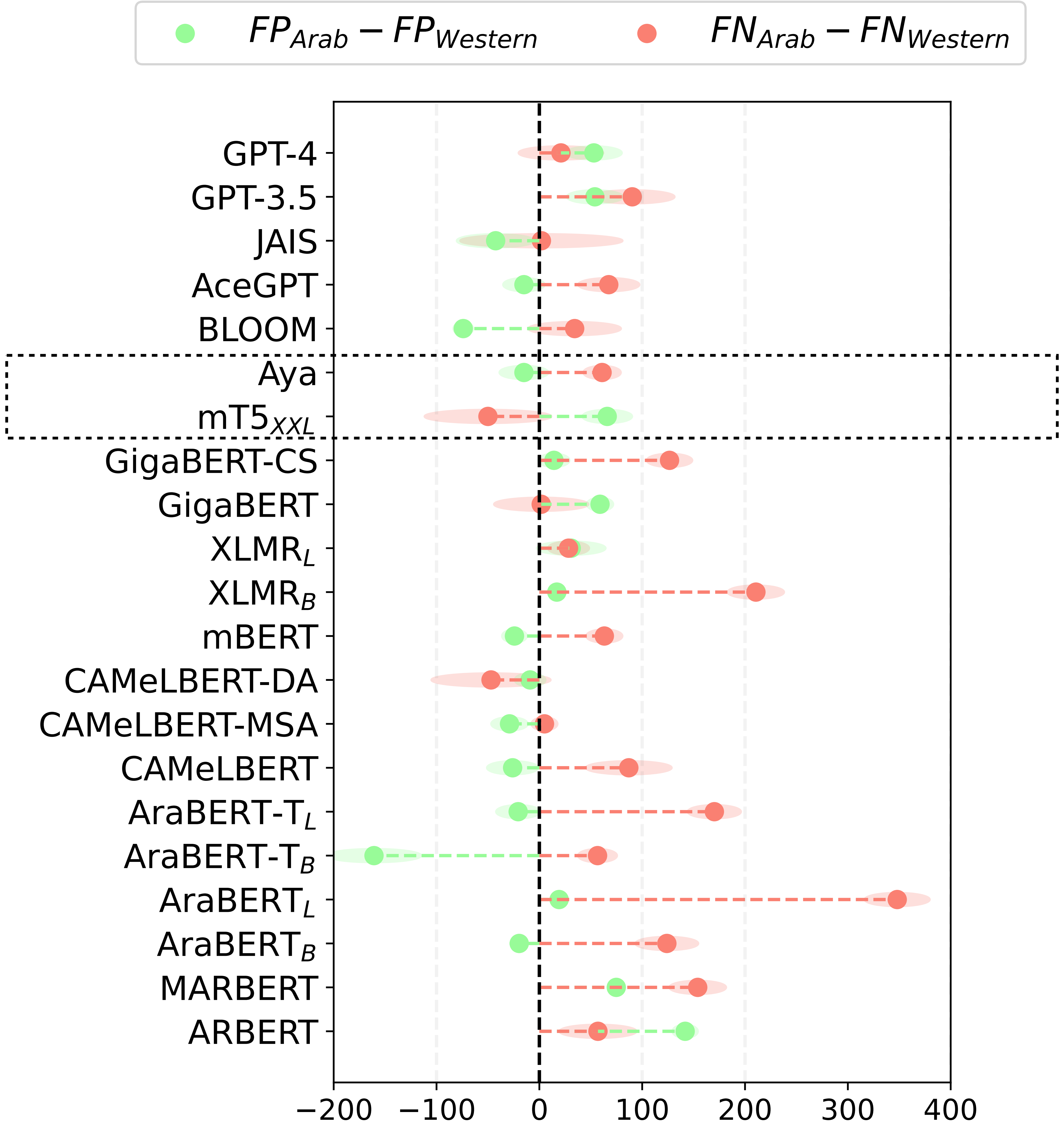
Note: CAMeL entities, prompts, and these adjectives are all in the Arabic language, but shown here in English on the slides for easy viewing.

CAMeL — What about Sentiment?



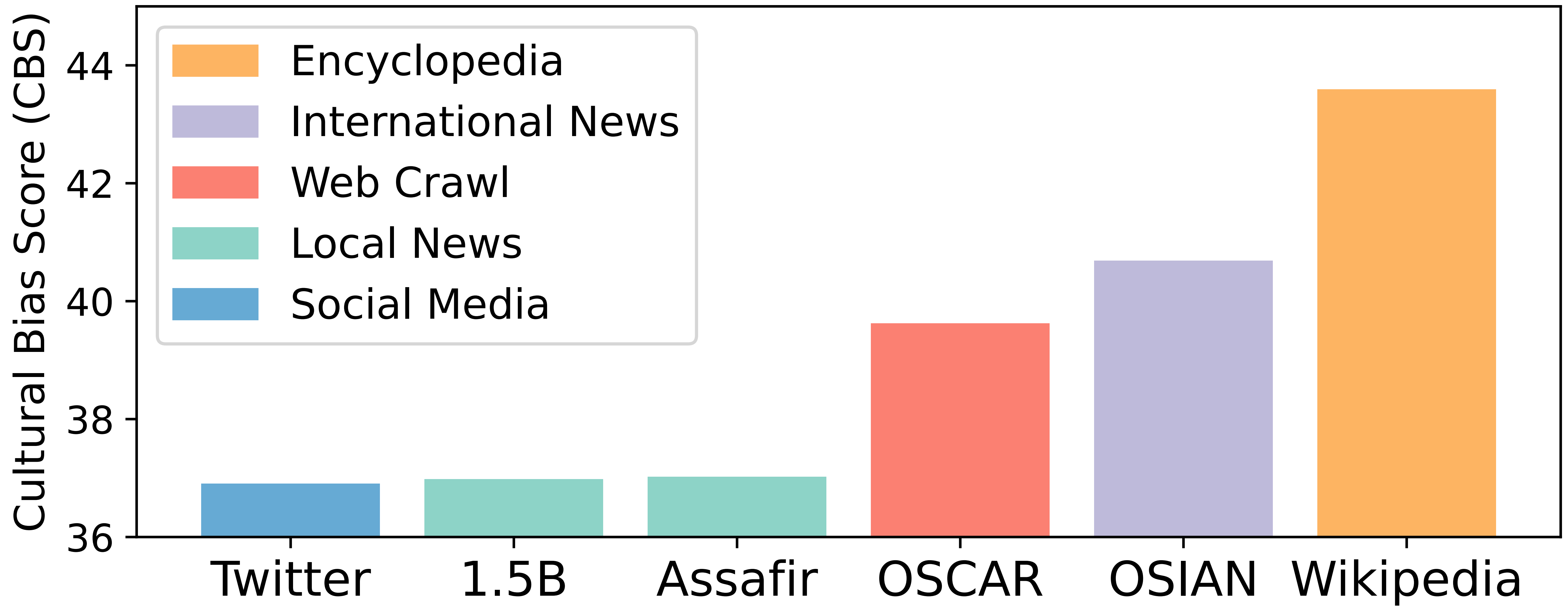
Note: CAMeL entities and prompts are all in the Arabic language, but shown here in English on the slides for easy viewing.

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)

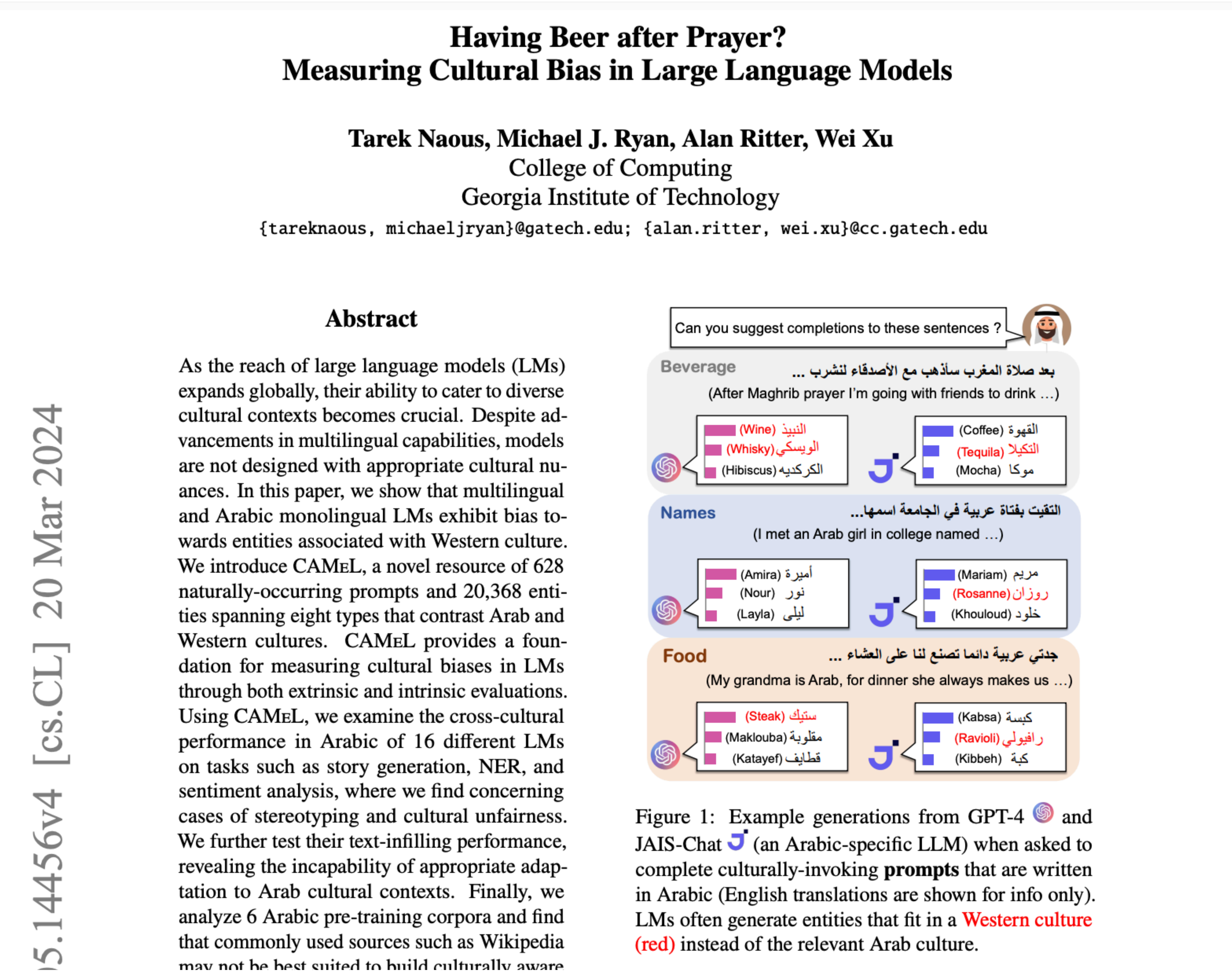


- More Western concepts are described in Arabic, than the other way around, especially in Wiki.
- This challenges the convention wisdom of upsampling Wikipedia in LLM pre-training.

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

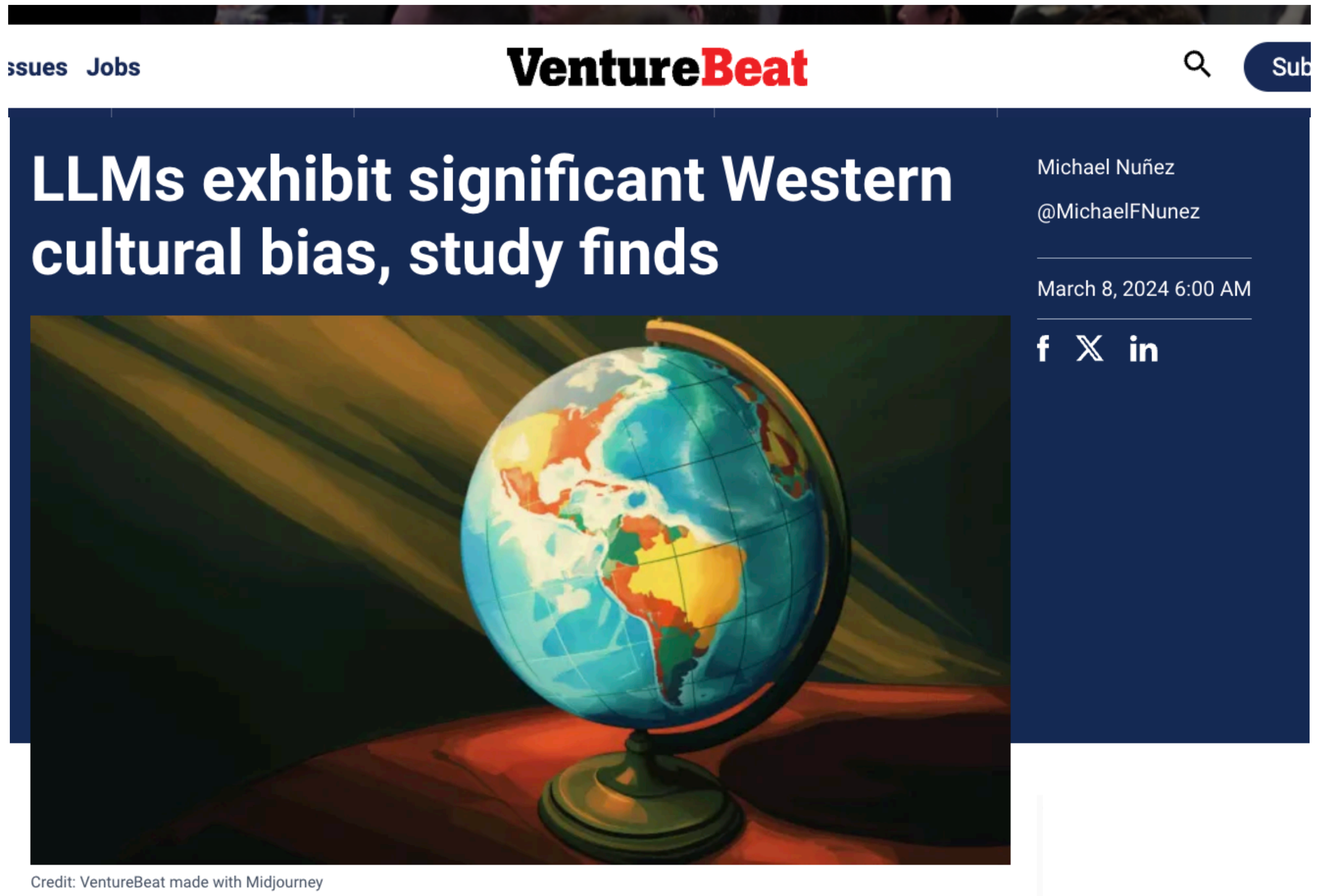
arXiv:2403.14456v4 [cs.CL] 20 Mar 2024

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



Tarek Naous, Michael J. Ryan, Alan Ritter, Wei Xu "Having Beer After Prayer? Measuring Cultural Bias in LLMs" (ACL 2024)

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

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

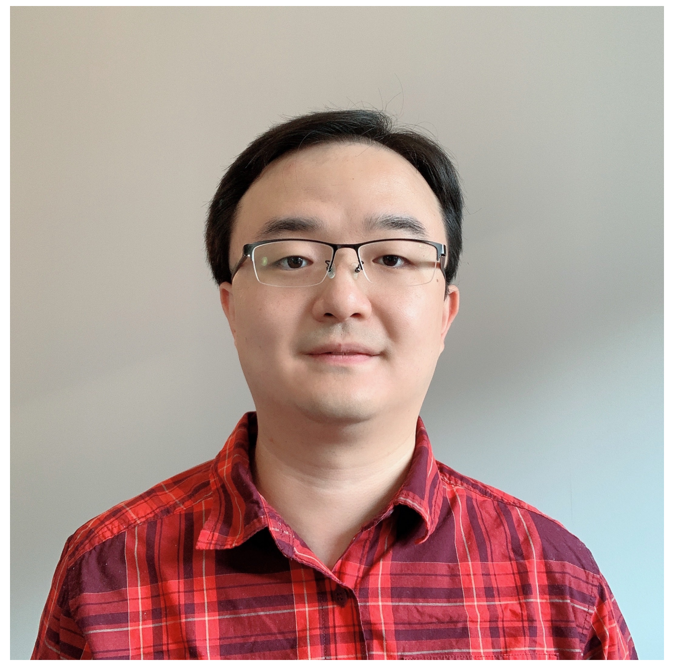
PrivacyMirror

(Yao et al., ACL 2024)

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

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

Yang Chen



Chao Jiang

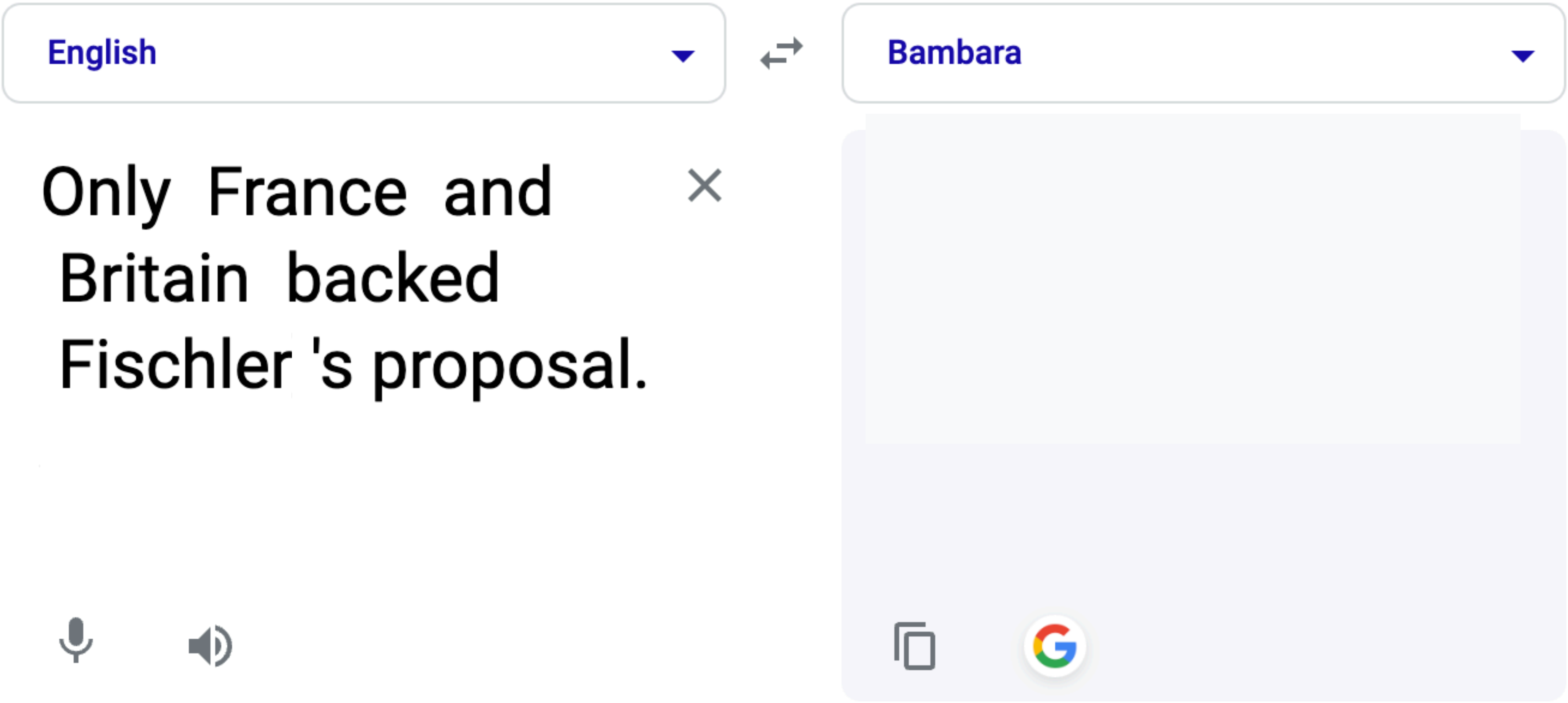
Alan Ritter

Wei Xu

A systematic study of marker-based approach for label projection

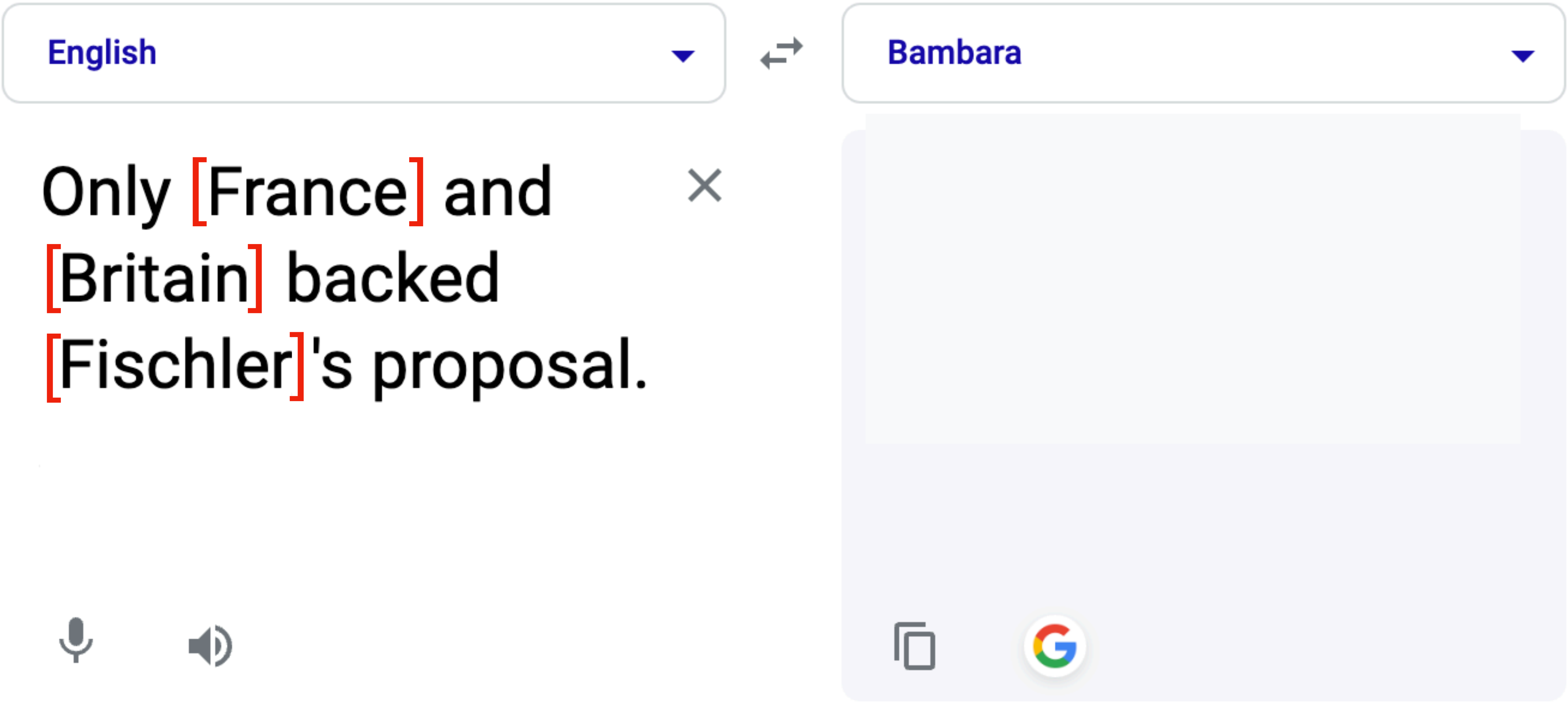
Marker-based Approach

Translating annotated training data from one language to the other



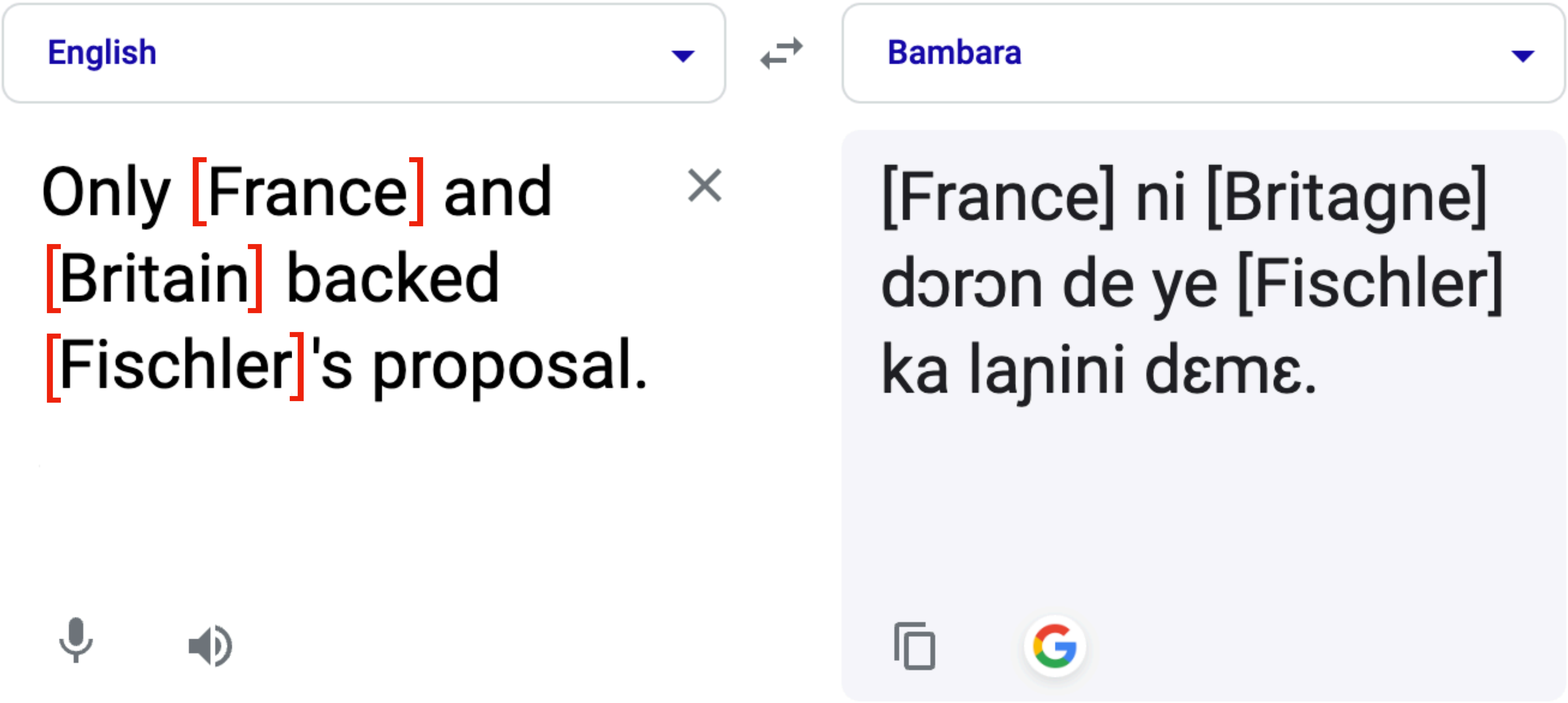
Marker-based Approach

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.



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.

EasyProject - Easy Marker-based Projection

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

works the best

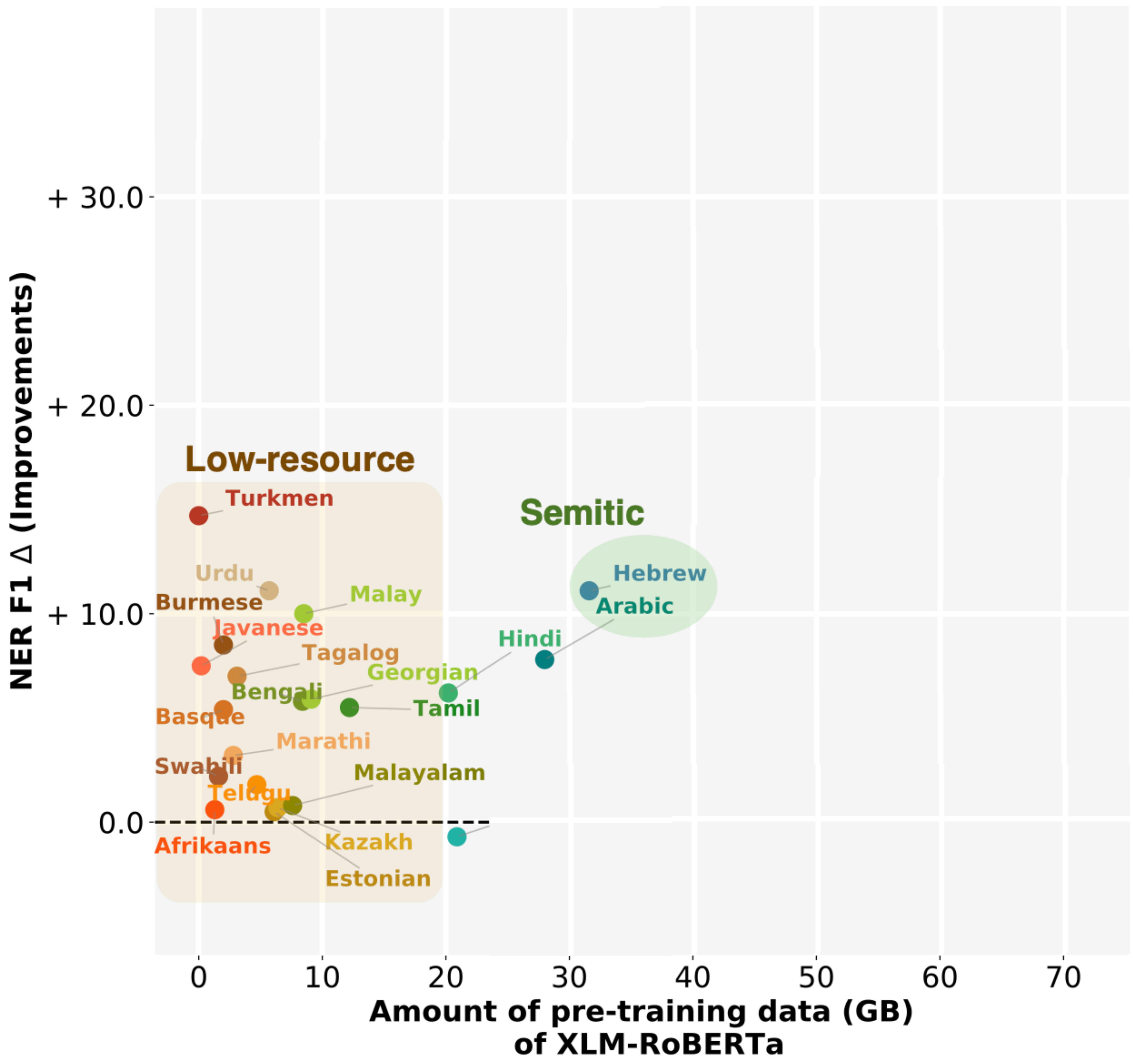
XML tags (e.g., <loc> </loc>) or [ ] “ ” ( ) < > { }

- 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

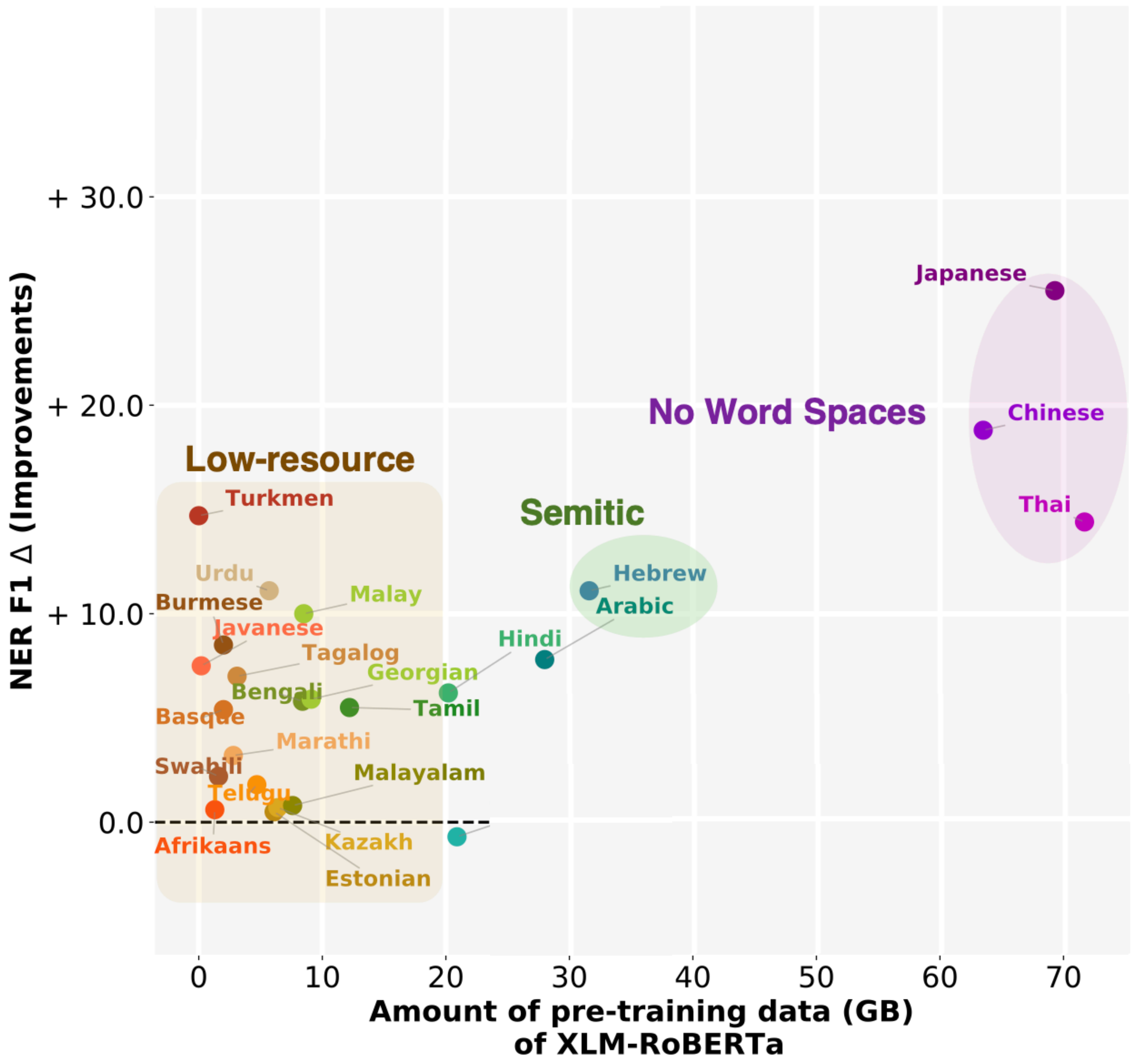
EasyProject - Easy Marker-based Projection

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



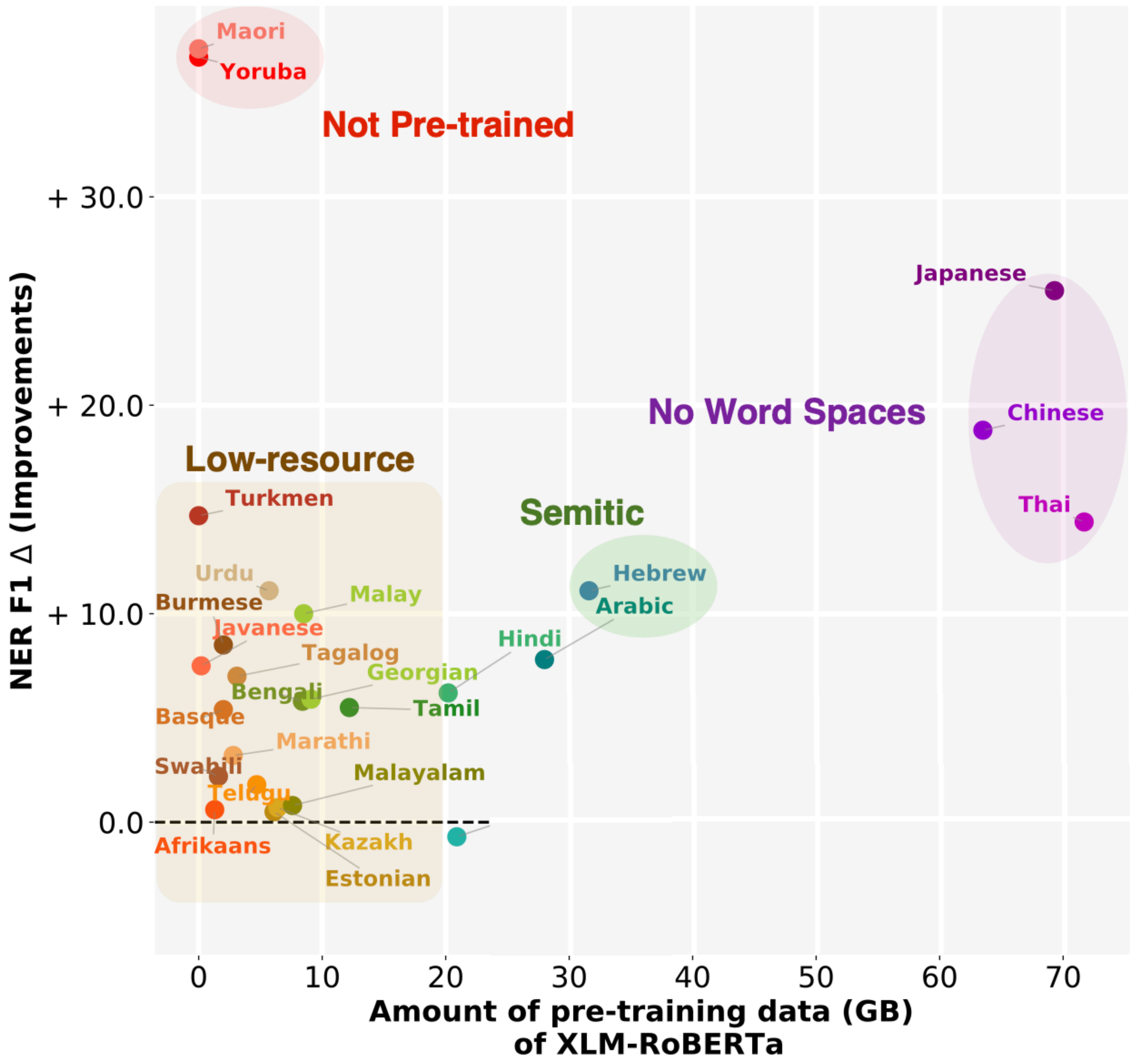
EasyProject - Easy Marker-based Projection

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



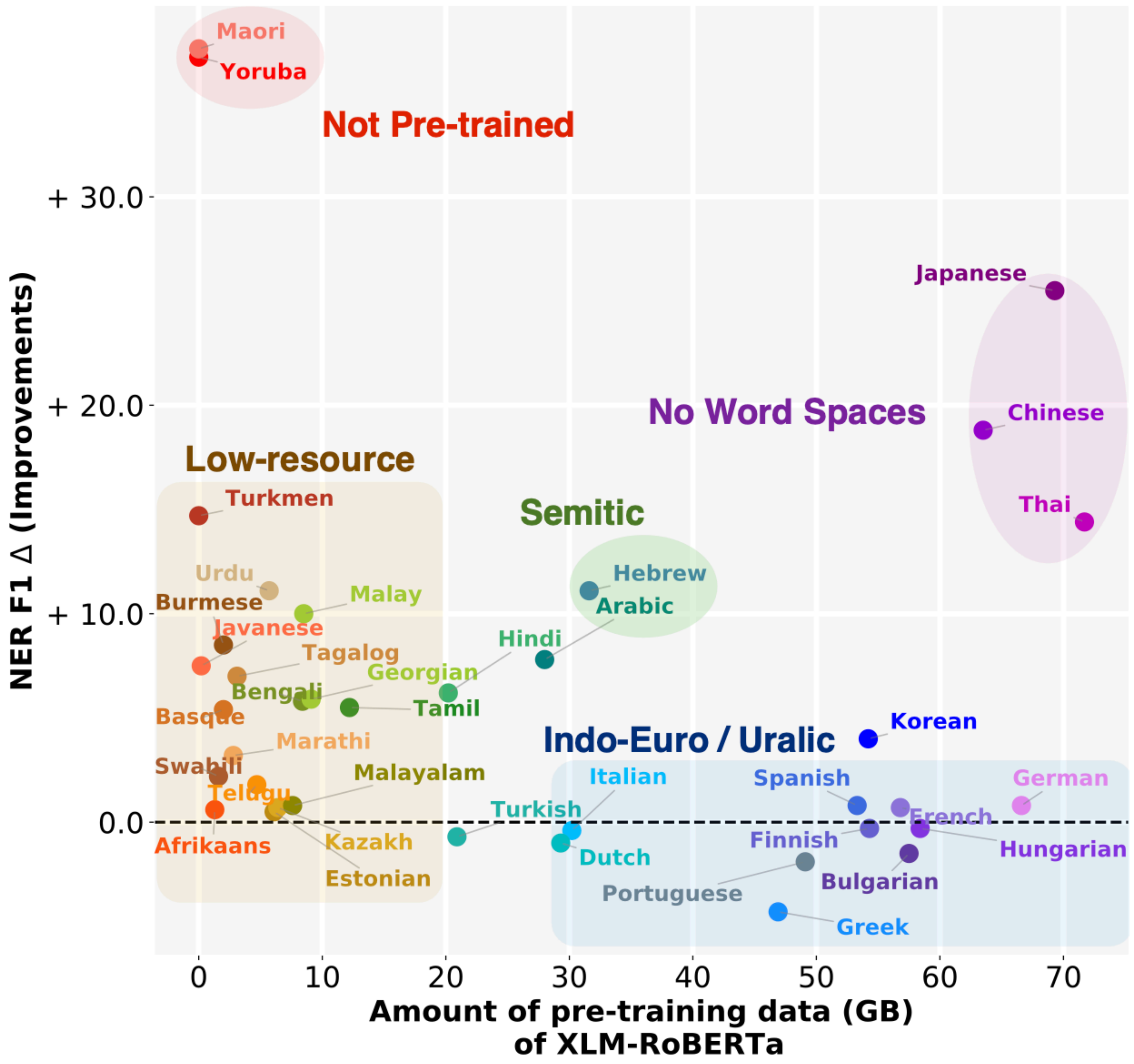
EasyProject - Easy Marker-based Projection

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



EasyProject - Easy Marker-based Projection

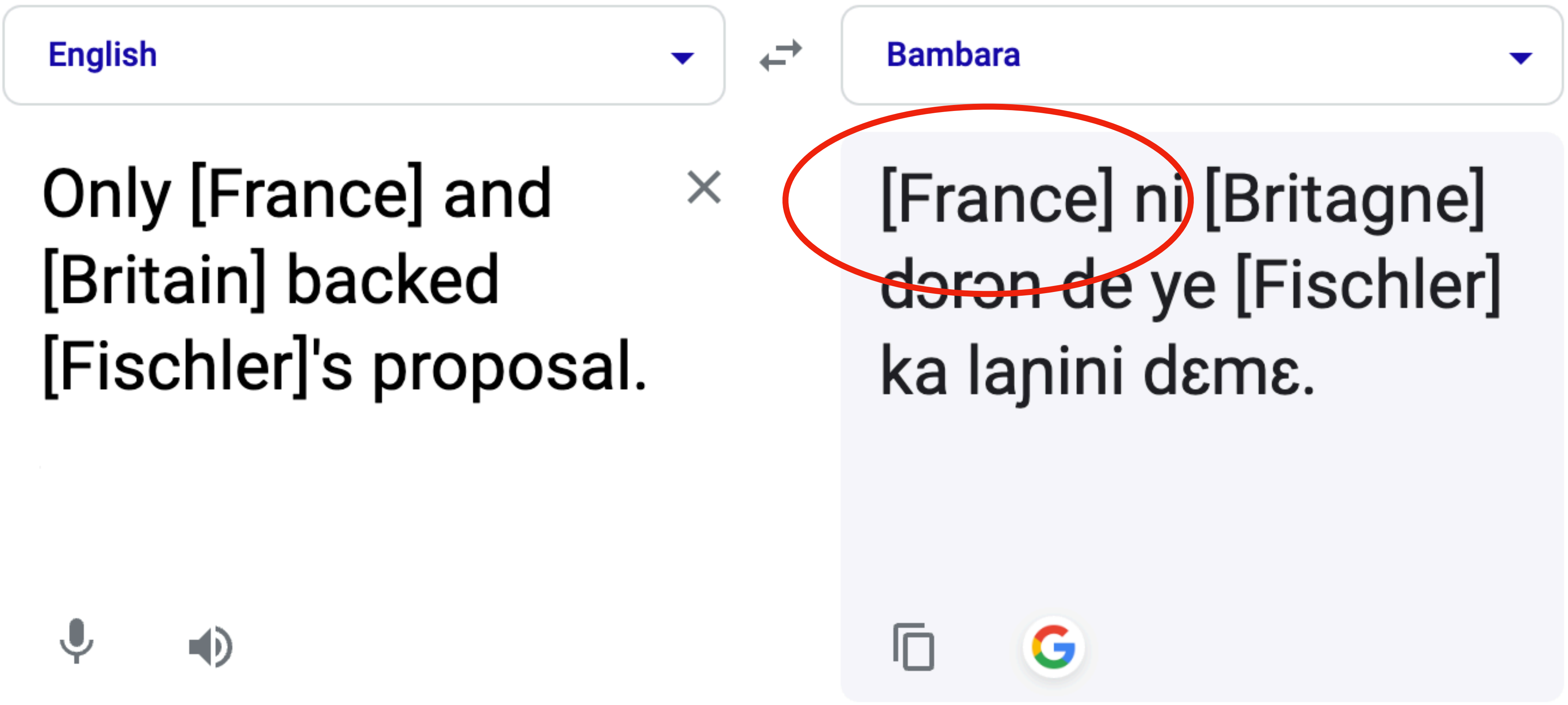
Especially promising for low-resource languages & languages that are written in non-Latin scripts



Zero-shot Cross-lingual Label Projection

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

marker-based approach



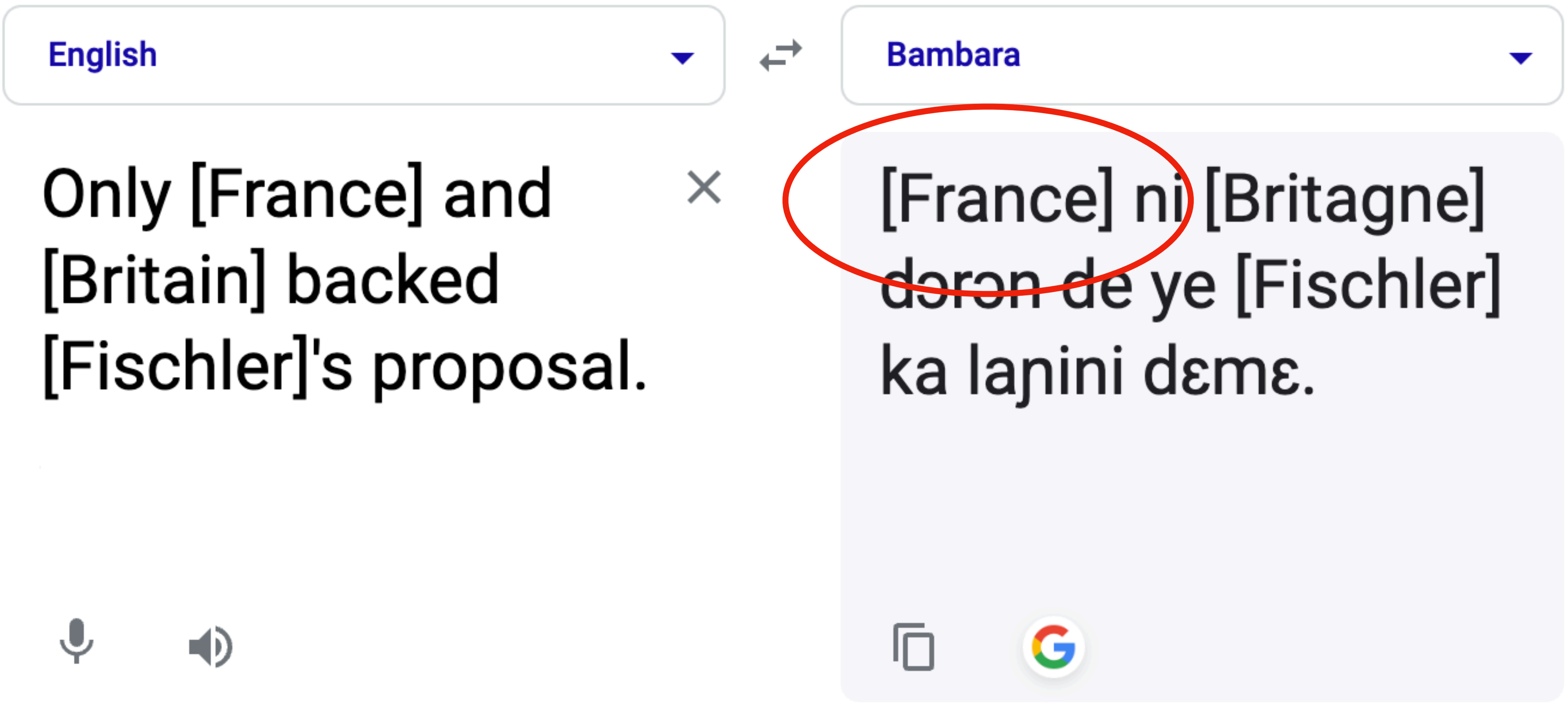
Only need a MT system
&
work surprisingly well !

But, degraded
MT quality
due to injected markers

Zero-shot Cross-lingual Label Projection

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

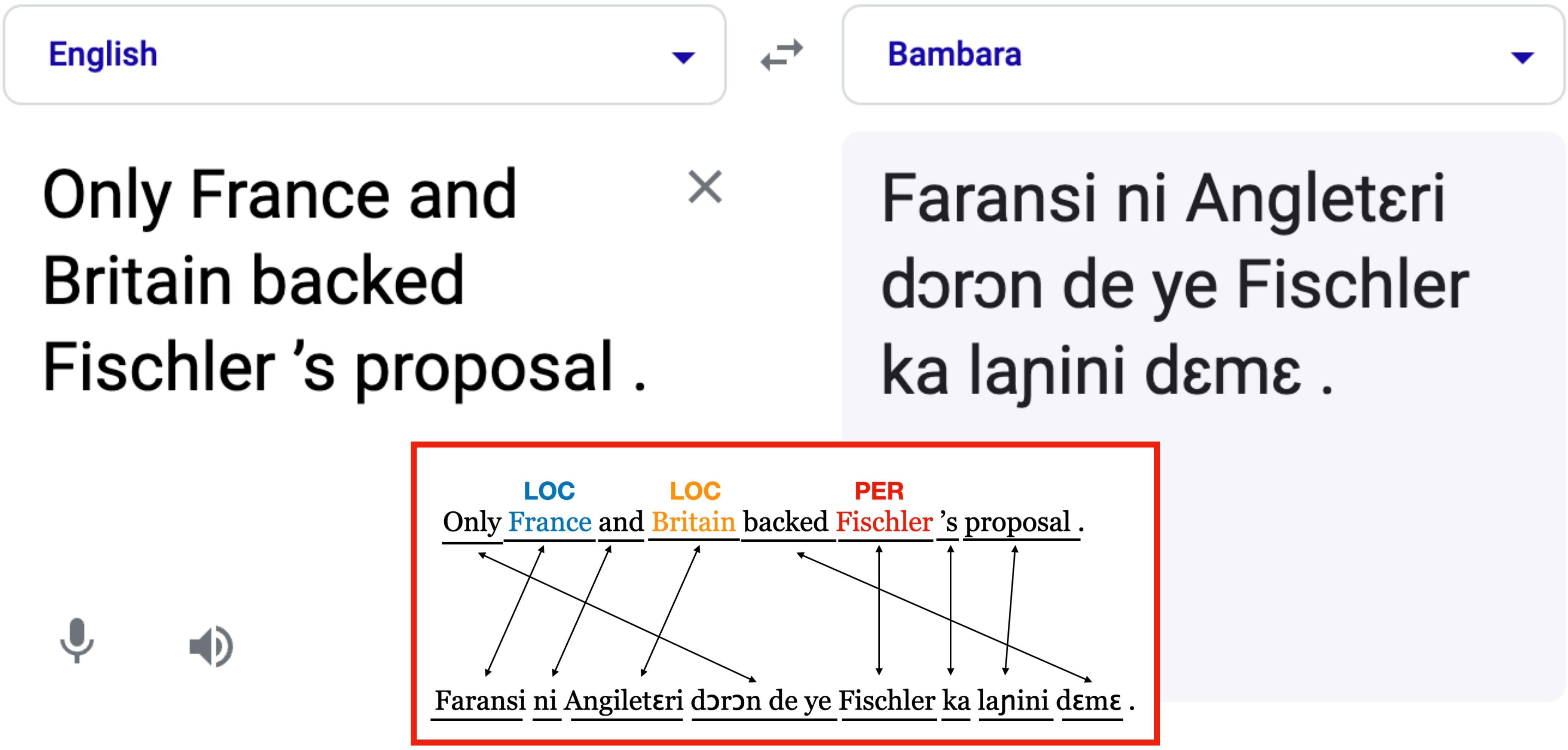
marker-based approach



Only need a MT system & work surprisingly well !

But, degraded MT quality due to injected markers

word alignment-based approach

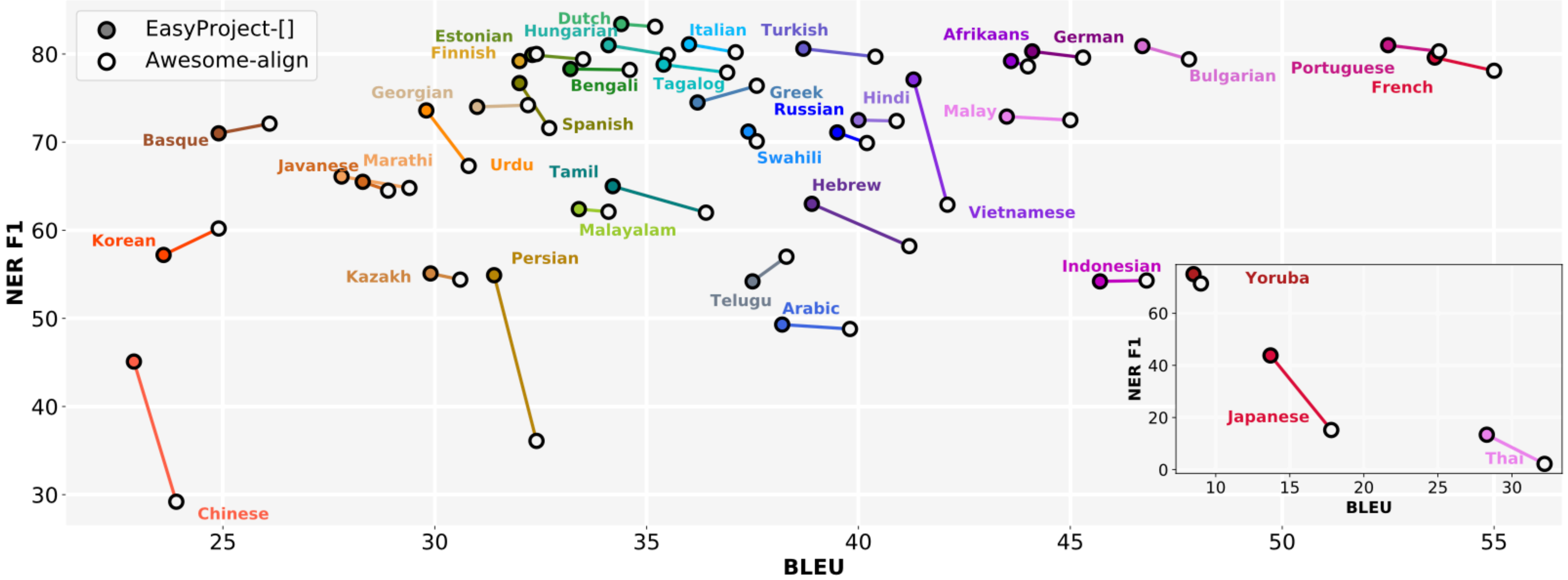


normally better MT quality

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!



Yang Chen, Chao Jiang, Alan Ritter, Wei Xu. “Frustratingly Easy Label Projection for Cross-lingual Transfer” (ACL 2023 Findings)

Can we do marker-based approach <u>without</u> scarifying the translation quality?

Constrained Decoding for Cross-lingual Label Projection (CODEC)

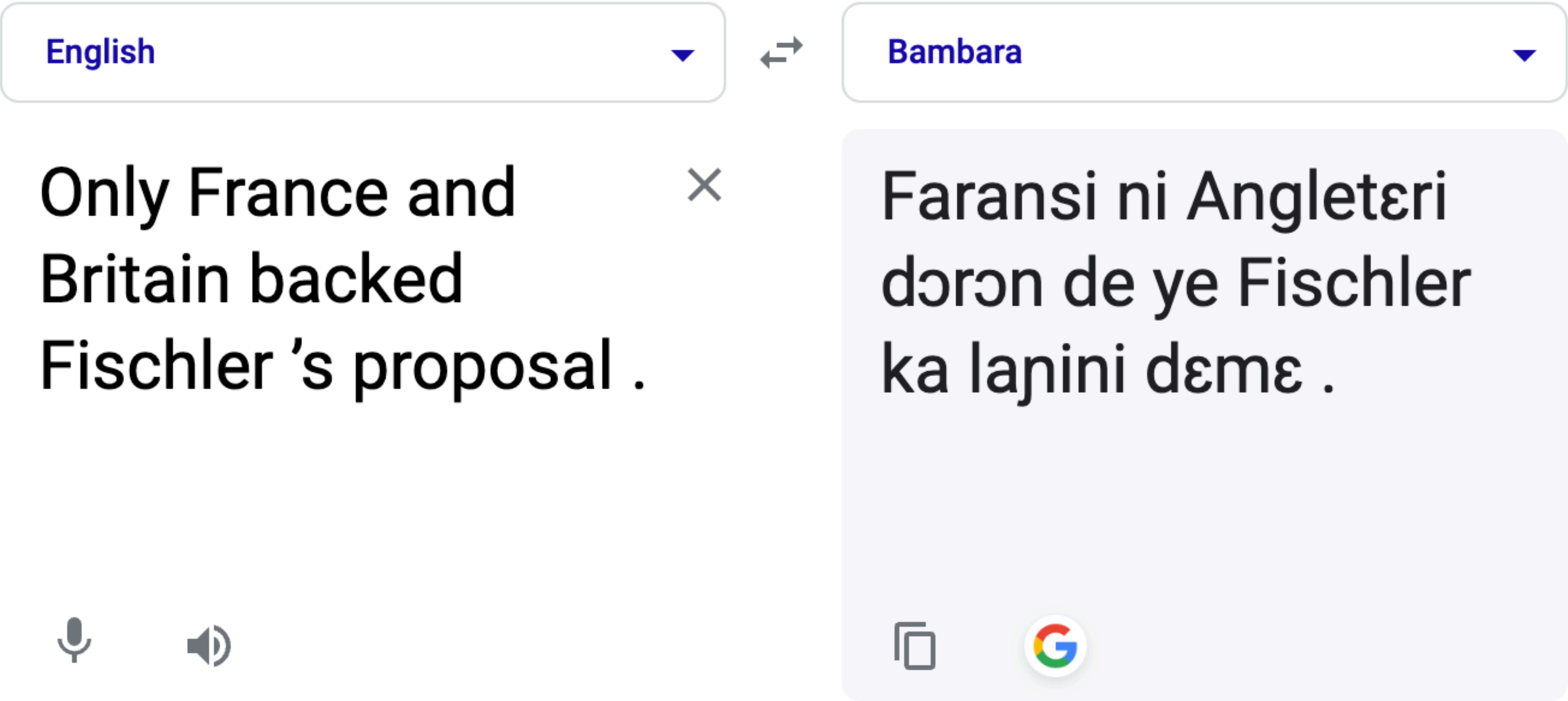



Duong Minh Le　　Yang Chen　　Alan Ritter　　Wei Xu

A better technical solution for marker-based label projection

Key Idea

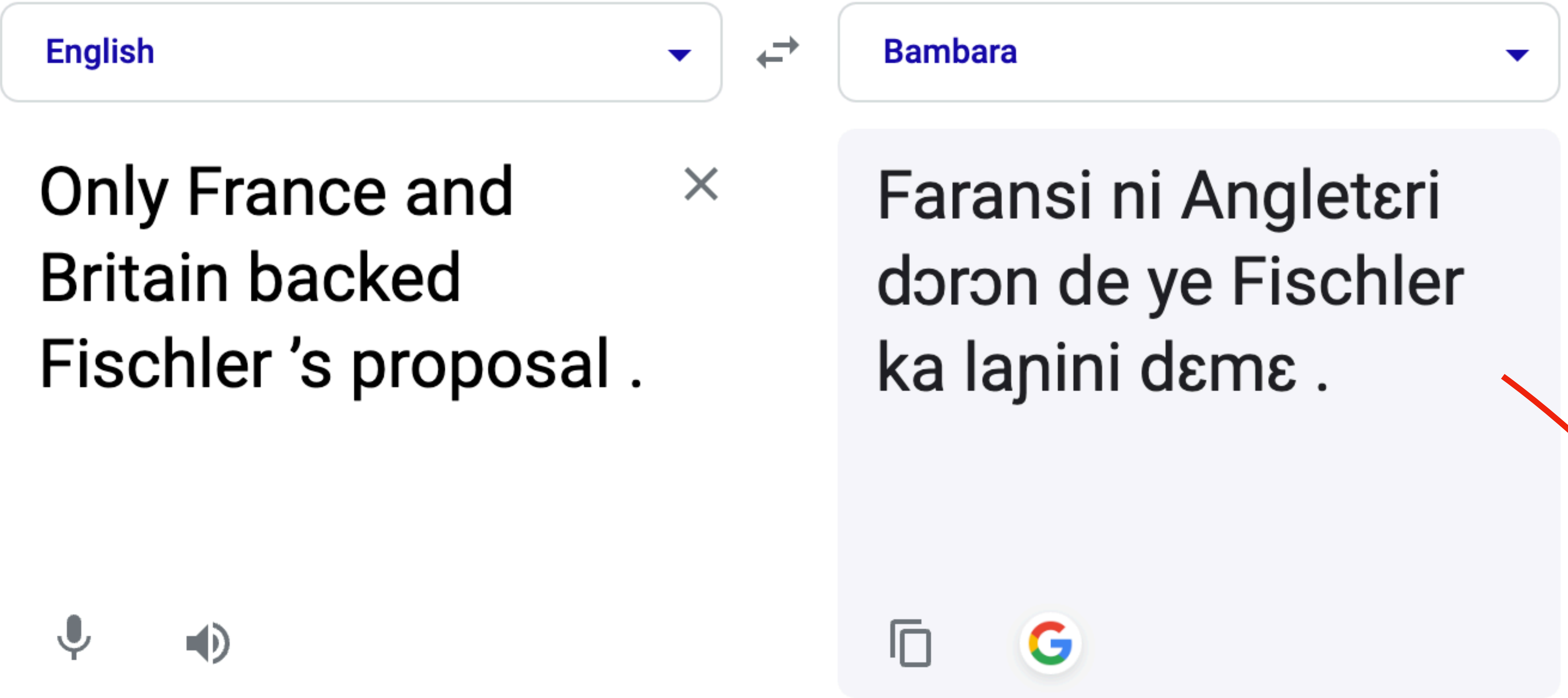
Step 1. Translate the original sentence as usual without markers.



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 sentece:

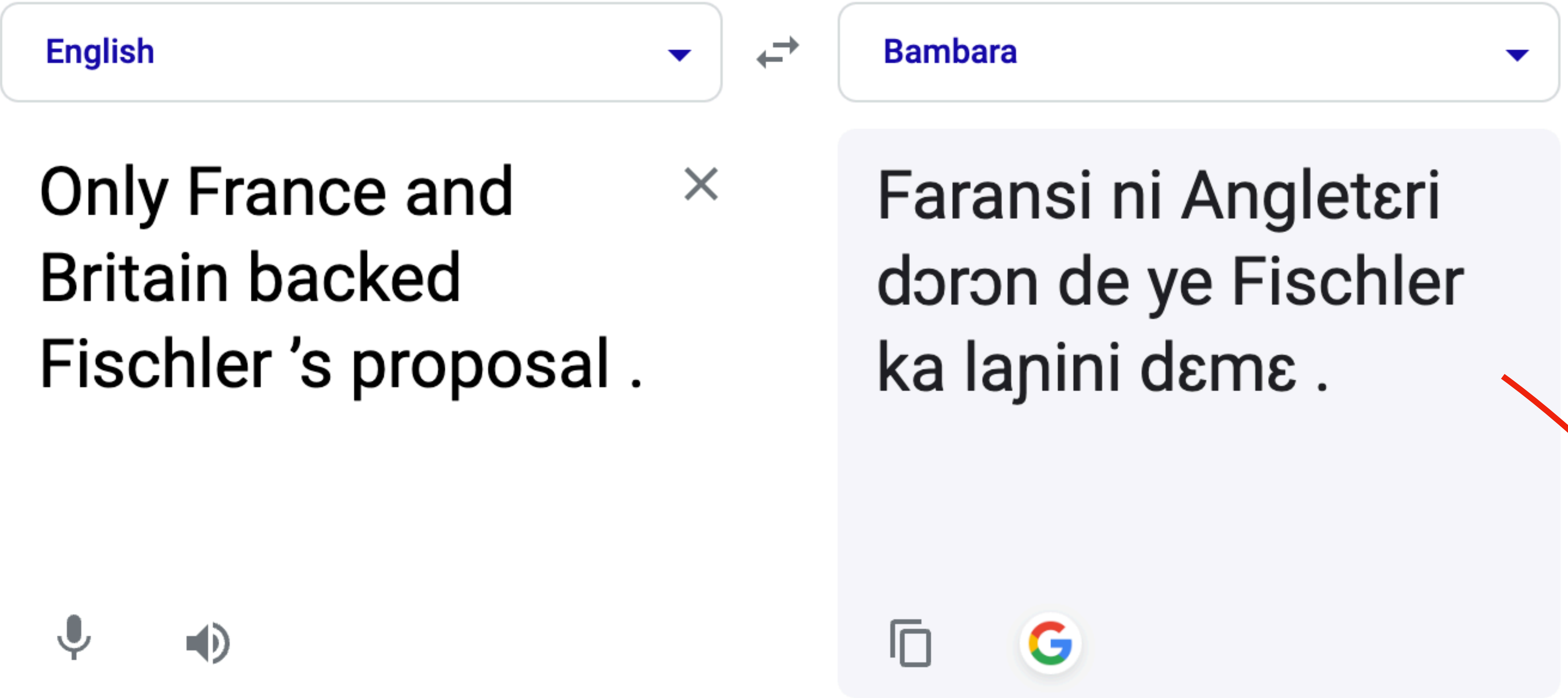
Only [France] and [Britain] backed [Fischler]'s proposal. →

LLMs
(Translation Models)

→ **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 sentece:

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

→

LLMs
(Translation Models)

→

Translated Output:

[Faransi] ni [Angiletɛri] dɔrɔn de ye [Fischler] ka laɲini 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} .

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

$$\boxed{O(n^{2m})}$$

An Efficient Constrained Decoding Algorithm

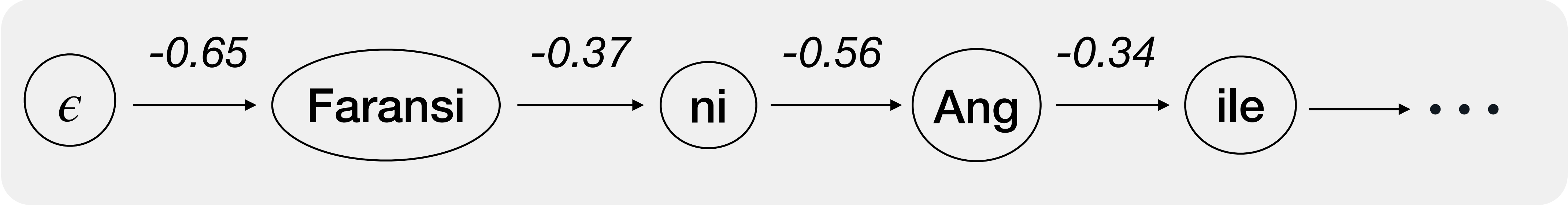
(1) Prune opening marker positions based on the contrastive log-likelihood difference.

An Efficient Constrained Decoding Algorithm

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

Input: $x =$ “Only France and Britain backed Fischler 's proposal .” $x^{mark} =$ “Only France and [Britain] backed Fischler 's proposal .” $y^{tmpl} =$ “Faransi ni Angiletɛri dɔrɔn de ye Fischler ka laɲini dɛmɛ .”

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



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

An Efficient Constrained Decoding Algorithm

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

Input: $x =$ "Only France and Britain backed Fischler 's proposal ." $x^{mark} =$ "Only France and [Britain] backed Fischler 's proposal ." $y^{tmpl} =$ "Faransi ni Angiletɛri dɔrɔn de ye Fischler ka laɲini dɛmɛ ."

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

$\epsilon \xrightarrow{-0.65}$ Faransi $\xrightarrow{-0.37}$ ni $\xrightarrow{-0.56}$ Ang $\xrightarrow{-0.34}$ ile $\longrightarrow \cdots$

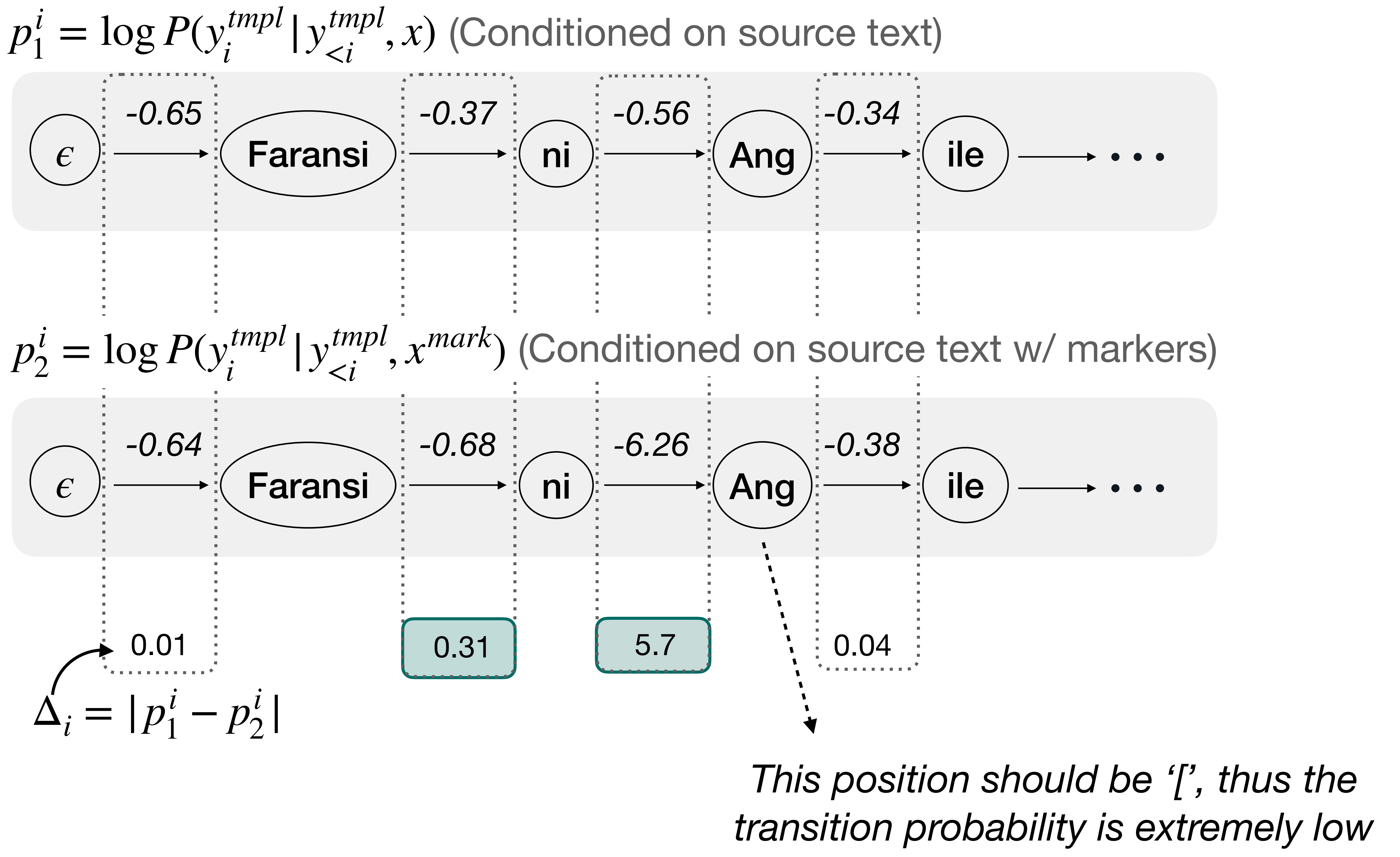
$p_2^i = \log P(y_i^{tmpl} | y_{<i}^{tmpl}, x^{mark})$ (Conditioned on source text w/ markers)

$\epsilon \xrightarrow{-0.64}$ Faransi $\xrightarrow{-0.68}$ ni $\xrightarrow{-6.26}$ Ang $\xrightarrow{-0.38}$ ile $\longrightarrow \cdots$

An Efficient Constrained Decoding Algorithm

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

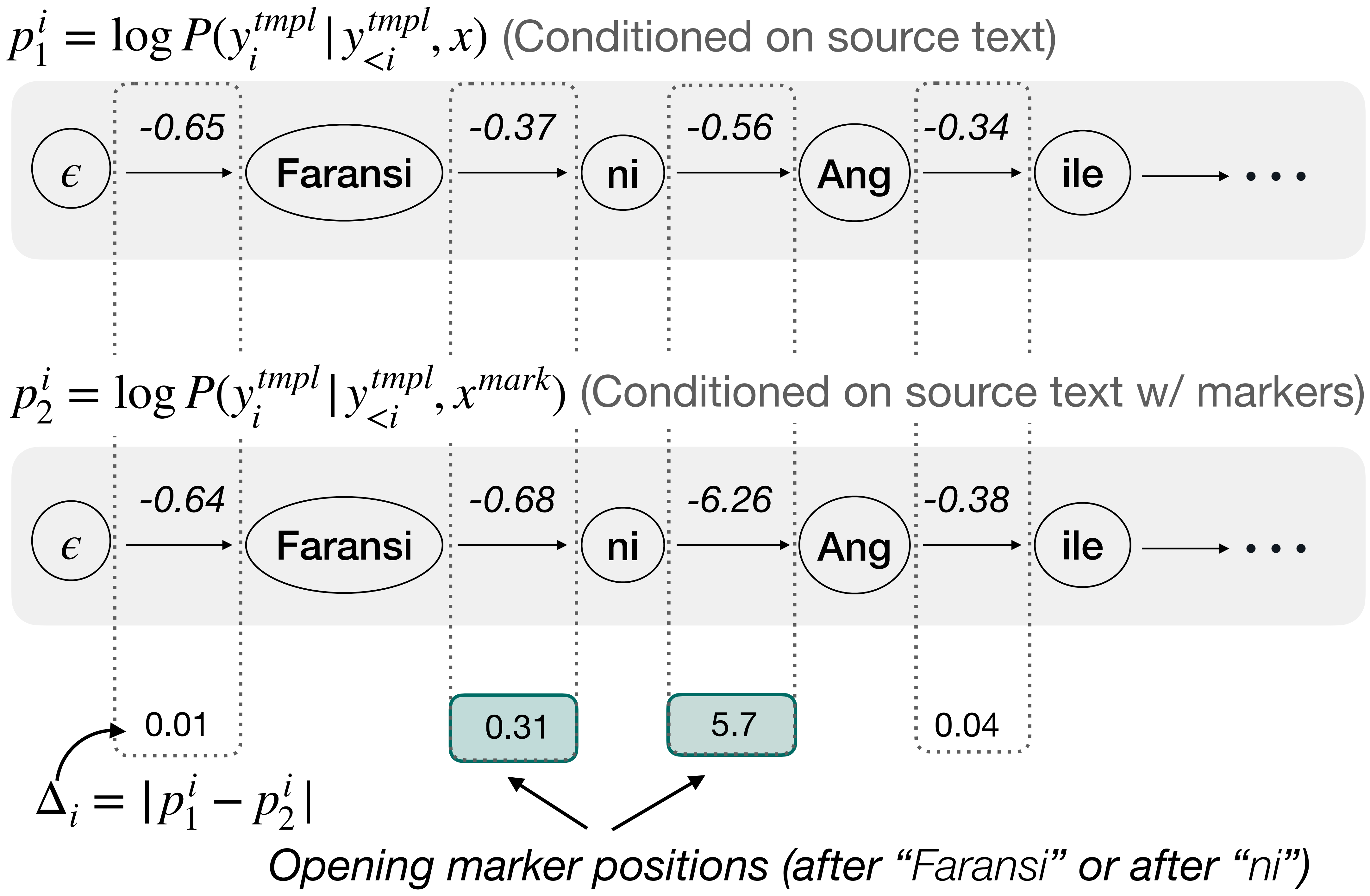
Input: $x =$ “Only France and Britain backed Fischler 's proposal .” $x^{mark} =$ “Only France and [Britain] backed Fischler 's proposal .” $y^{tmpl} =$ “Faransi ni Angiletɛri dɔrɔn de ye Fischler ka laɲini dɛmɛ .”



An Efficient Constrained Decoding Algorithm

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

Input: $x =$ “Only France and Britain backed Fischler 's proposal .” $x^{mark} =$ “Only France and [Britain] backed Fischler 's proposal .” $y^{tmpl} =$ “Faransi ni Angiletɛri dɔrɔn de ye Fischler ka laɲini dɛmɛ .”



An Efficient Constrained Decoding Algorithm

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

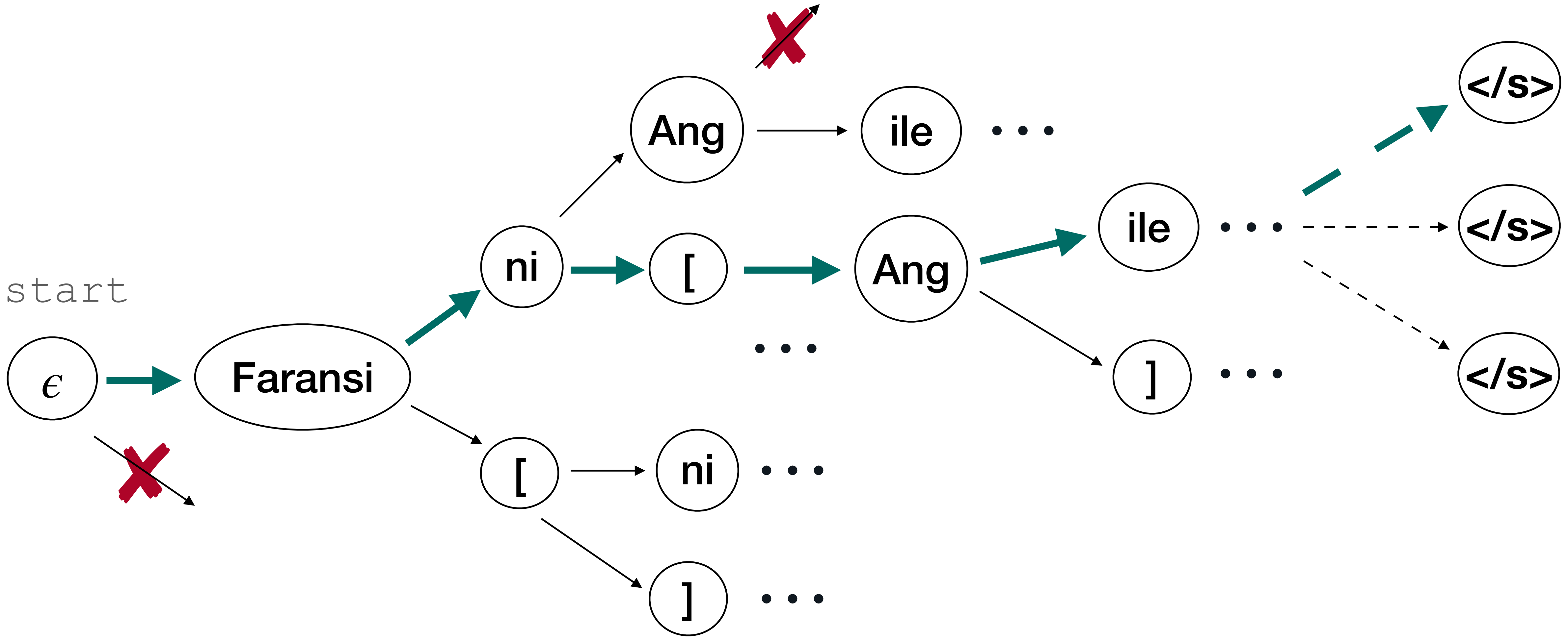
$$d = \min\left(\max\left(j + \delta, q\right), |y^k|\right)$$

An Efficient Constrained Decoding Algorithm

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

$$d = \min\left(\max\left(j+\delta, q\right), |y^k|\right)$$

Input: $x =$ "Only France and Britain backed Fischler 's proposal ." $x^{mark} =$ "Only France and [Britain] backed Fischler 's proposal ." $y^{tmpl} =$ "Faransi ni Angilɛtɛri dɔrɔn de ye Fischler ka laɲini dɛmɛ ."



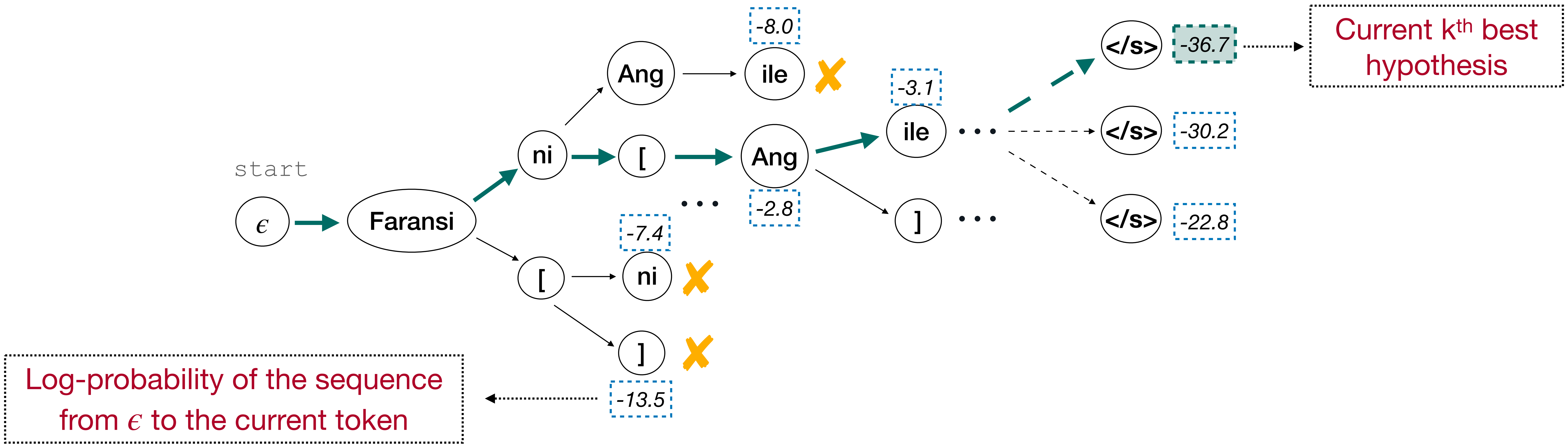
Prune opening-marker positions

An Efficient Constrained Decoding Algorithm

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

$$d = \min\left(\max\left(j + \delta, q\right), |y^k|\right)$$

Input: $x =$ "Only France and Britain backed Fischler 's proposal ." $x^{mark} =$ "Only France and [Britain] backed Fischler 's proposal ." $y^{tmpl} =$ "Faransi ni Angiletɛri dɔrɔn de ye Fischler ka laɲini dɛmɛ ."

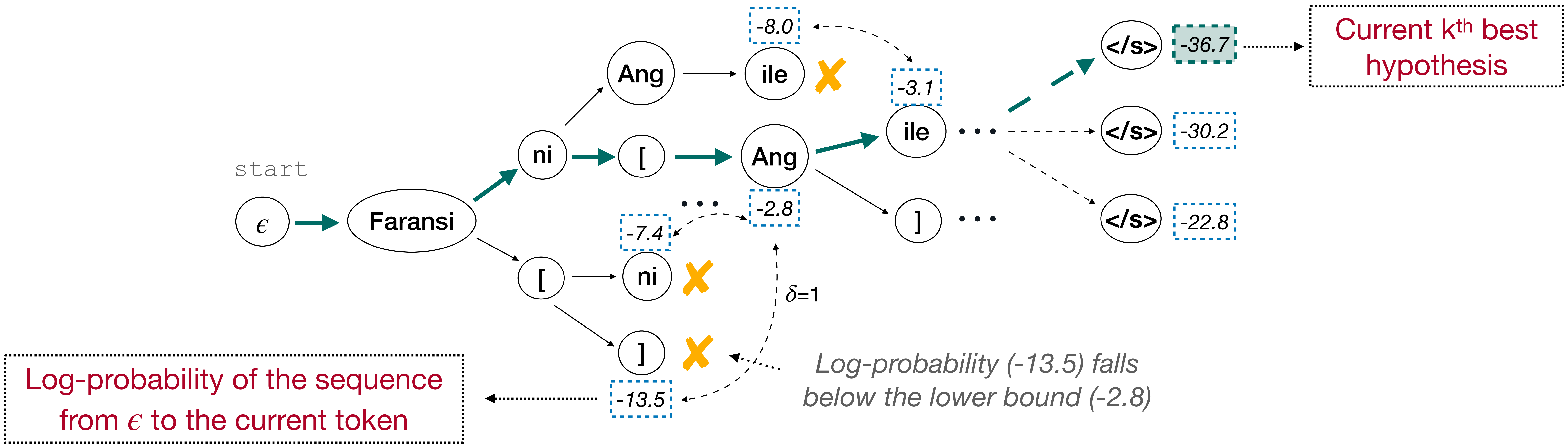


An Efficient Constrained Decoding Algorithm

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

$$d = \min\left(\max\left(j + \delta, q\right), |y^k|\right)$$

Input: $x =$ “Only France and Britain backed Fischler 's proposal .” $x^{mark} =$ ”Only France and [Britain] backed Fischler 's proposal .” $y^{tmpl} =$ ”Faransi ni Angileteri dɔrɔn de ye Fischler ka laɲini dɛmɛ .”



Prune branches based on a heuristic lower-bound

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: $[\log P(y_1|x), \log P(y_{1:2}|x), \ldots, \log P(y|x)]$ (default=[0.0]), $\mathcal{M}$: opening marker positions
H: min heap to record the results, k: number of hypotheses, δ: lower bound hyperparameter

$flag \leftarrow$ {check if all markers are generated}
if $y_{|y|} = \texttt{</s>}$ and $flag = \texttt{TRUE}$: **then**
    $H.\mathrm{push}((L_{|y|}, L, y))$ ▷ H sorts by the first element
    if $\mathrm{len}(H) > k$ **then**
        $H.\mathrm{pop}()$
else
    $\mathcal{T} \leftarrow []$
    $w_1 \leftarrow$ {get the next token in y^{tmpl}}
    $\mathcal{T} \leftarrow \mathcal{T} \cup \{(w_1, \log P(w_1|y, x^{mark}))\}$
    $j \leftarrow |y| + 1$ ▷ position of the token to be generated next
    $w_2 \leftarrow$ {get the next marker}
    if $\exists\, w_2$ and not ($w_2 =$ '[' land $j \notin \mathcal{M}$) **then**
        $\mathcal{T} \leftarrow \mathcal{T} \cup \{(w_2, \log P(w_2|y, x^{mark}))\}$
    $\mathcal{T} \leftarrow$ {sort $\mathcal{T}$ by the second element in decreasing order}
    for $(w, p) \in \mathcal{T}$ **do**
        $logp \leftarrow L_{|y|} + p$
        $\gamma \leftarrow$ {compute lower bound following Eq 7}
        if $logp > \gamma$ **then**
            Constrained_DFS$(x^{mark}, y \cdot w, y^{tmpl}, L \cup \{logp\}, \mathcal{M}, H, k, \delta)$
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-align*[1]) 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

[1]*Zi-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

Experiment Results

prior marker-based approach cannot do this

"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.

only marks sub-words as an entity

	English Data	Augmented data in low-resource languages: EasyProject	Awesome-align	Codec
chiShona	India$_{LOC}$ and Pakistan$_{LOC}$ have fought … region of Kashmir$_{LOC}$ …	India$_{LOC}$ ne Pakistan$_{LOC}$ … ye Kashmir$_{LOC}$ chibviro …	India$_{LOC}$ nePakistan … zvinetso yeKashmir$_{LOC}$ …	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 … we China$_{LOC}$ ne Taipei$_{LOC}$, uTang Shubei$_{PER}$, … elivela eTaiwan$_{LOC}$ …	Imithombo$_{LOC}$ … waseChina neTaipei , uTang Shubei$_{PER}$, … elivela eTaiwan …	Imithombo … waseChina$_{LOC}$ neTaipei$_{LOC}$, uTang Shubei$_{PER}$, … elivela eTaiwan$_{LOC}$ …

having difficulty to project multiple spans

Duong Minh Le, Yang Chen, Alan Ritter, Wei Xu. "Constrained Decoding for Cross-lingual Label Projection" (ICLR 2024)

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

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.

A user-study informed NLP model design

Reducing Privacy Risks in Online Self-Disclosures (PrivacyMirror)


Yao Dou


Isadora Krsek


Tarek Naous


Anubha Kabra


Sauvik Das


Alan Ritter


Wei Xu

Georgia Tech

Carnegie Mellon University

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 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

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 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 ···

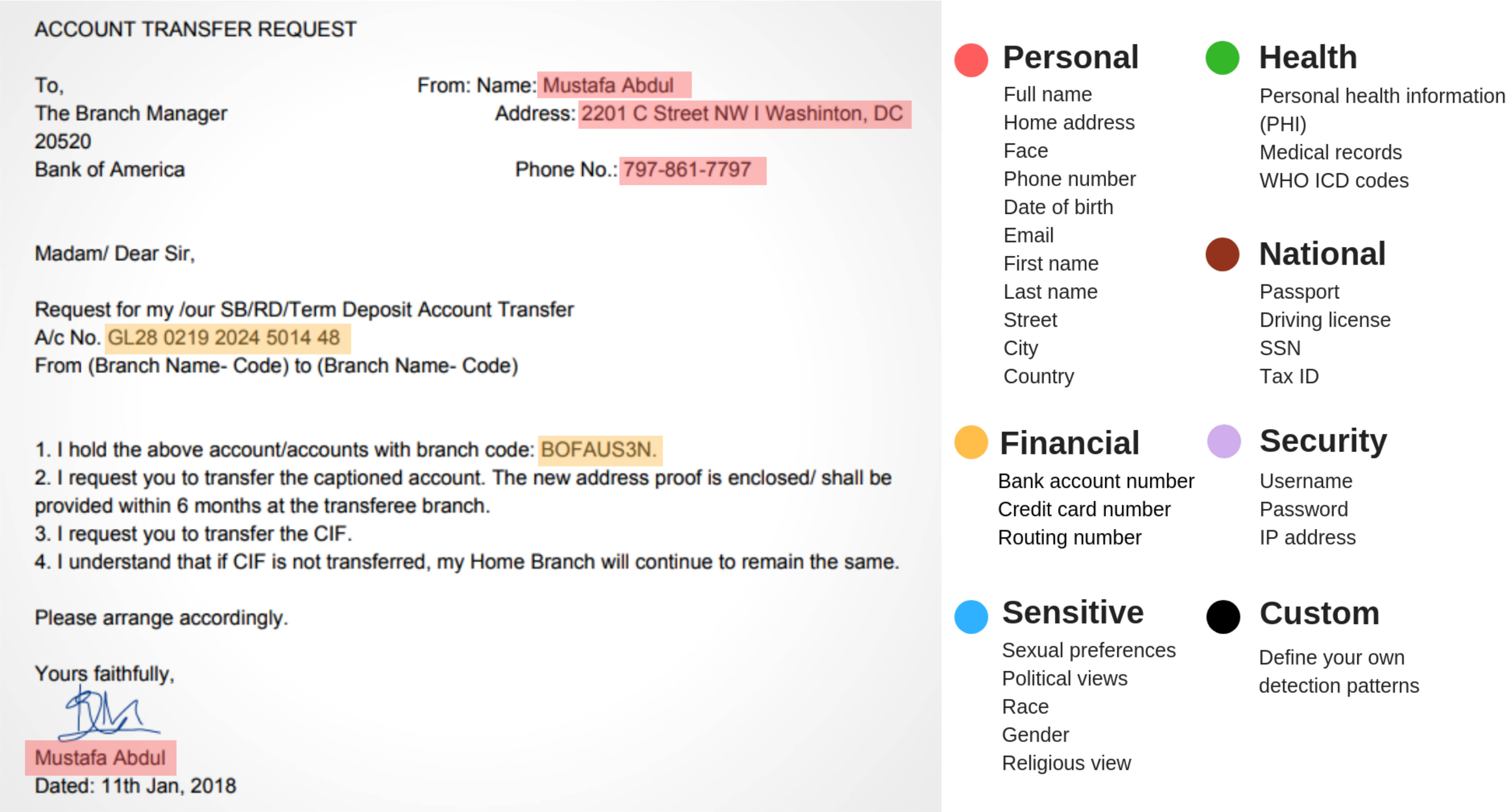
Disclosures:

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 xxxyyyzzz@gmail.com"

🪞 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

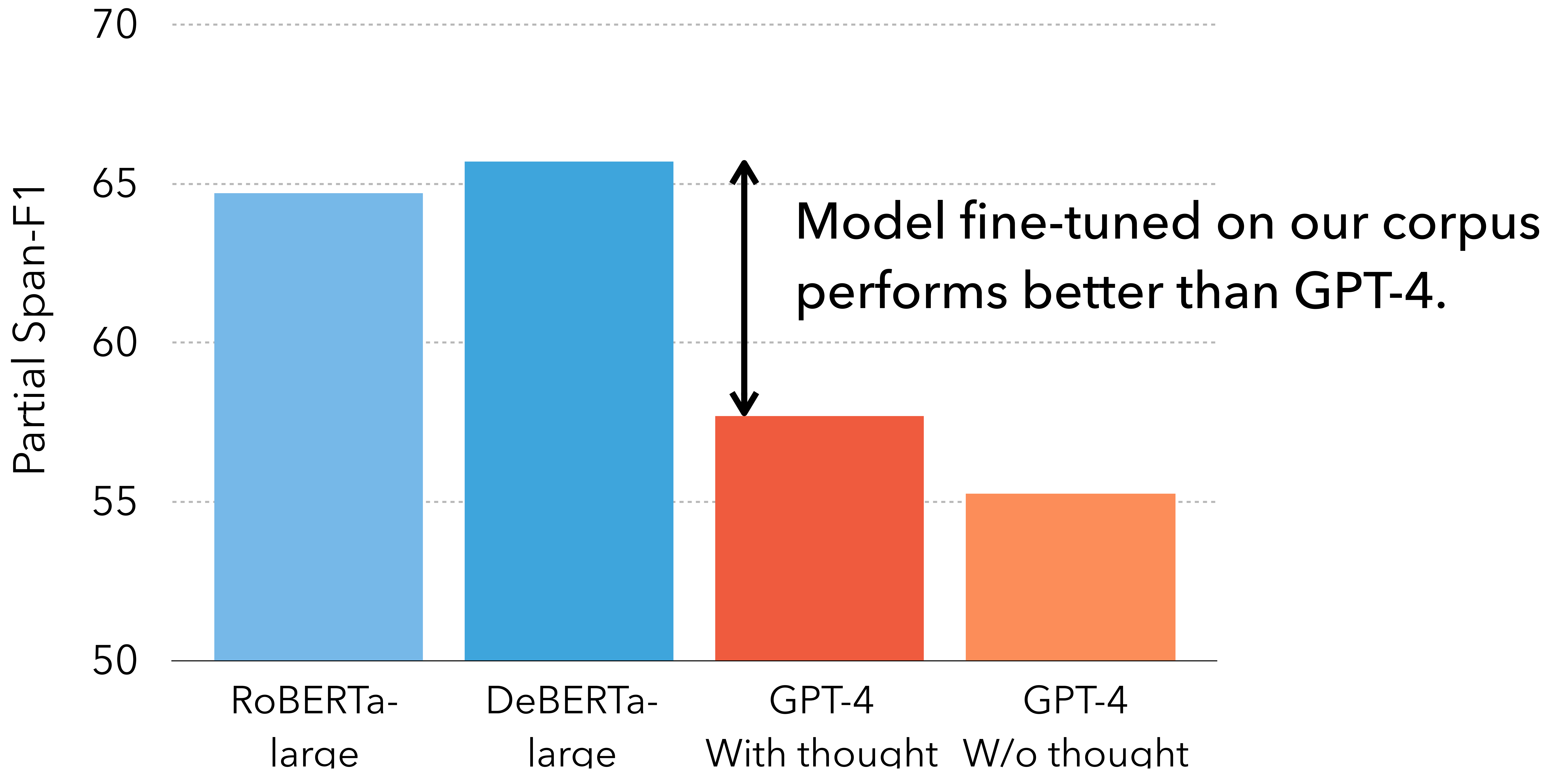
Hi there, I got accepted to UCLA (IS), which I'm pumped about.

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

My husband and I vote for different parties

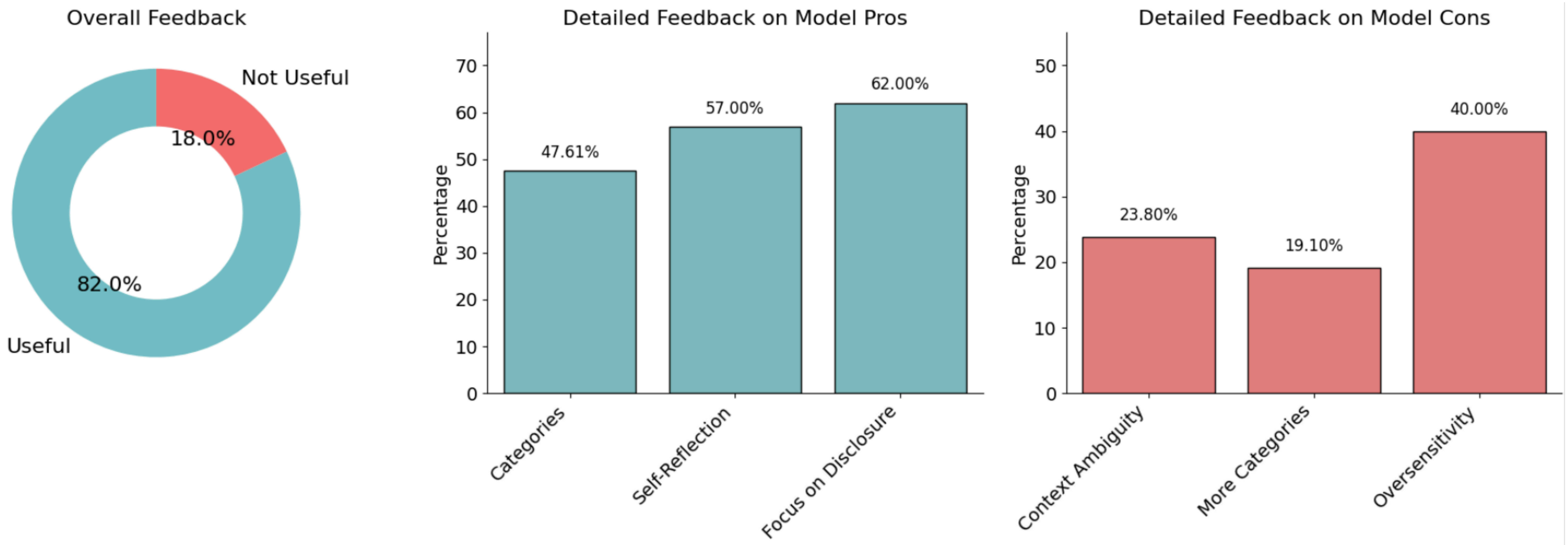
PrivacyMirror — Self-disclosure Detection

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



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 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.

✗ Utility

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

✗ Privacy

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.

✗ 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

✗ Privacy

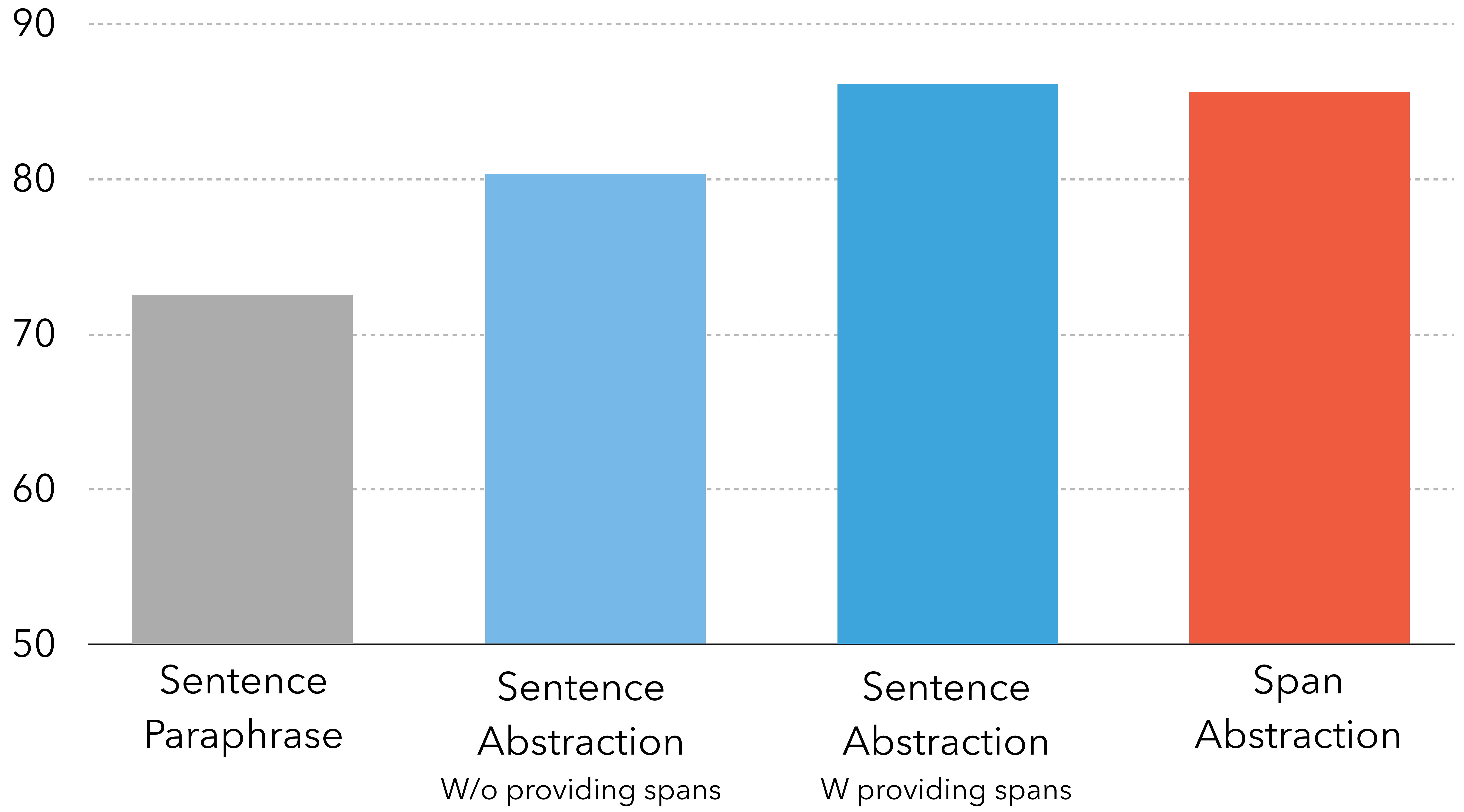
Sentence Parh I’m in Korea, I can’t actually drink because I’m not 21 yet.

Sentence Abough to legally consume alcohol even though I’m abroad.

✗ Writing Style

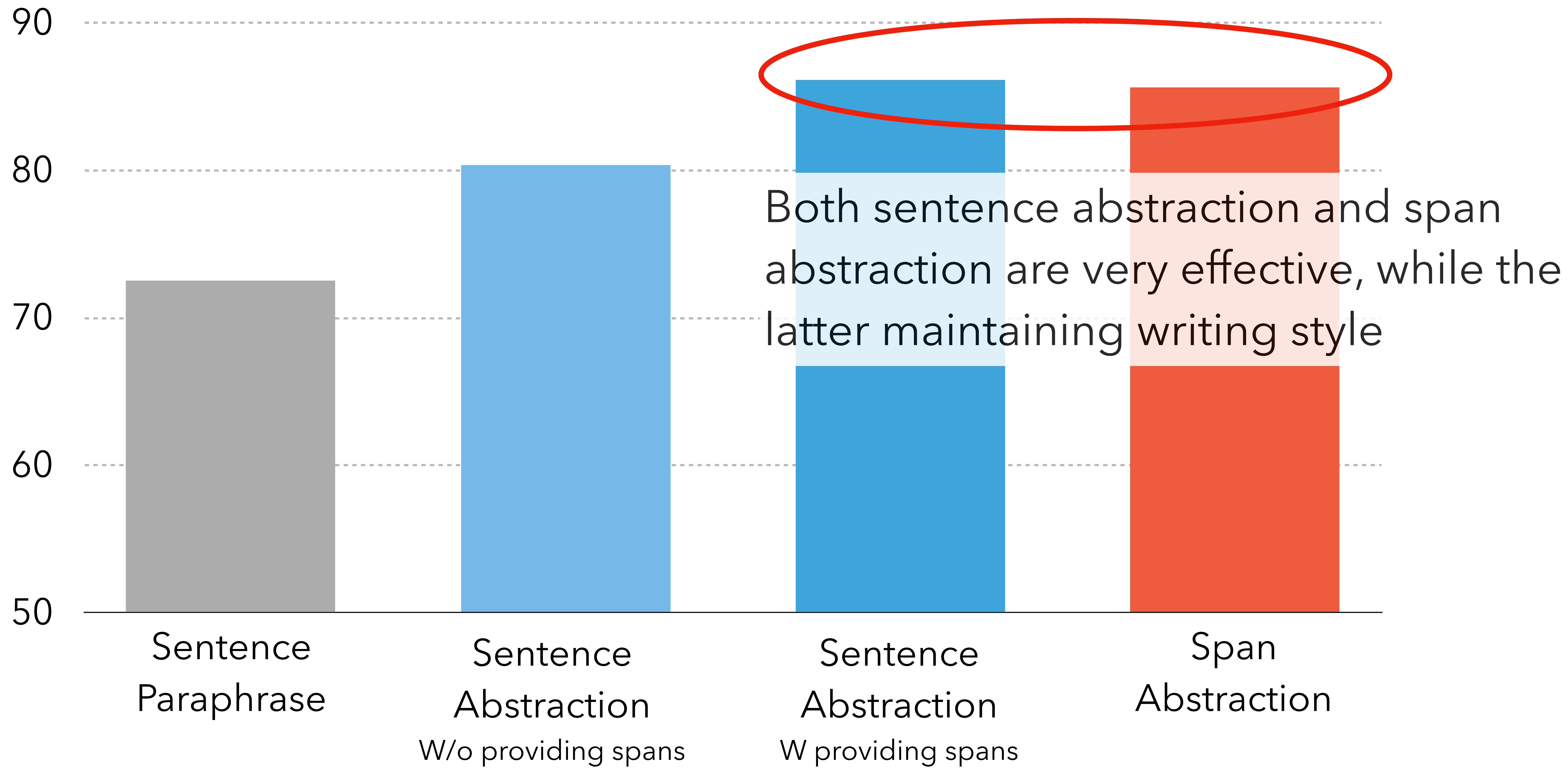
PrivacyMirror — Self-disclosure Abstraction

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

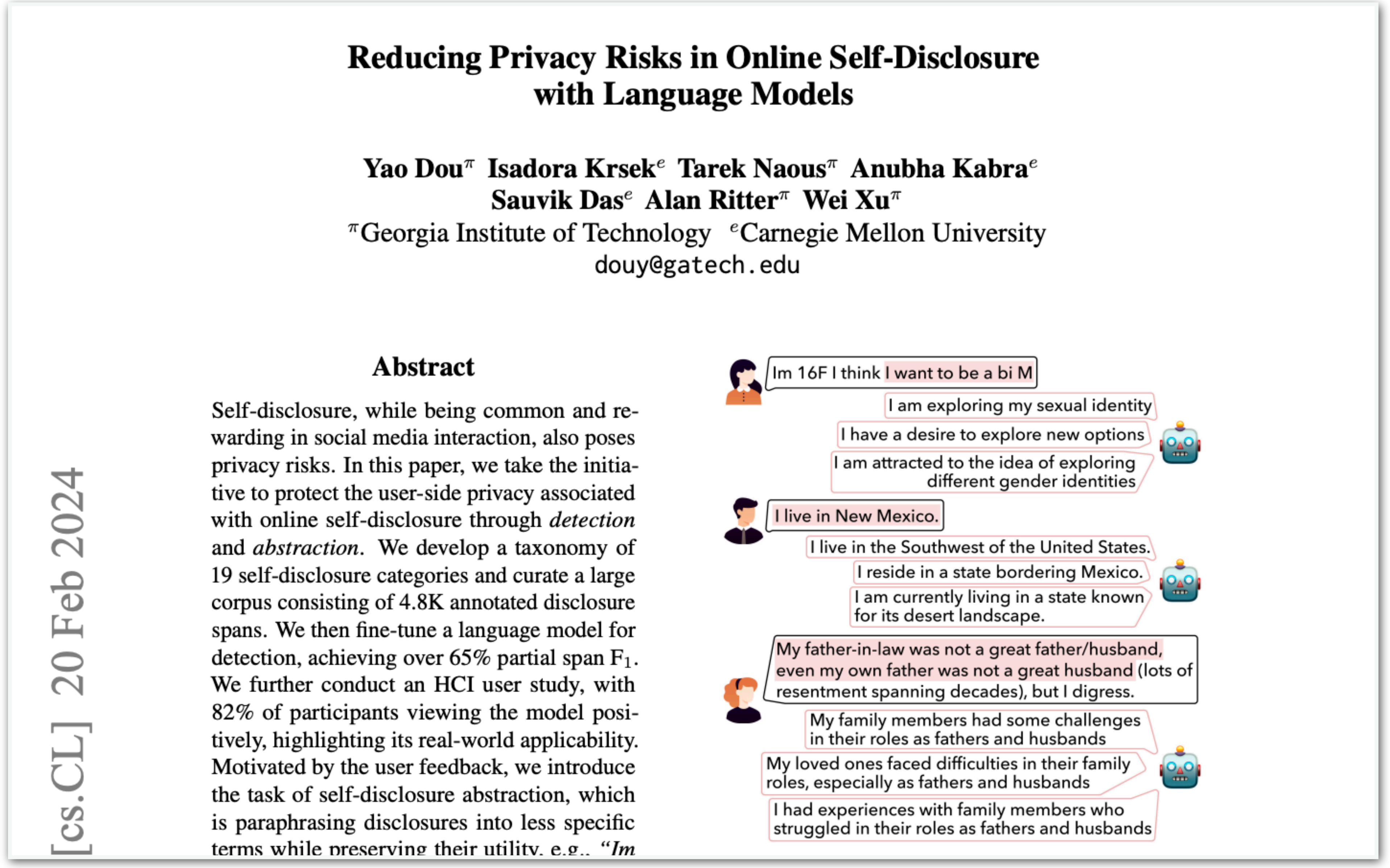


PrivacyMirror — Takeaways

- **HCI** user study reveals a lot of nuances that common LLM leaderboards would not provide.

- 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



Reducing Privacy Risks in Online Self-Disclosure with Language Models

Yao Dou[π] Isadora Krsek[c] Tarek Naous[π] Anubha Kabra[c]
Sauvik Das[c] Alan Ritter[π] Wei Xu[π]
[π]Georgia Institute of Technology [c]Carnegie Mellon University
douy@gatech.edu

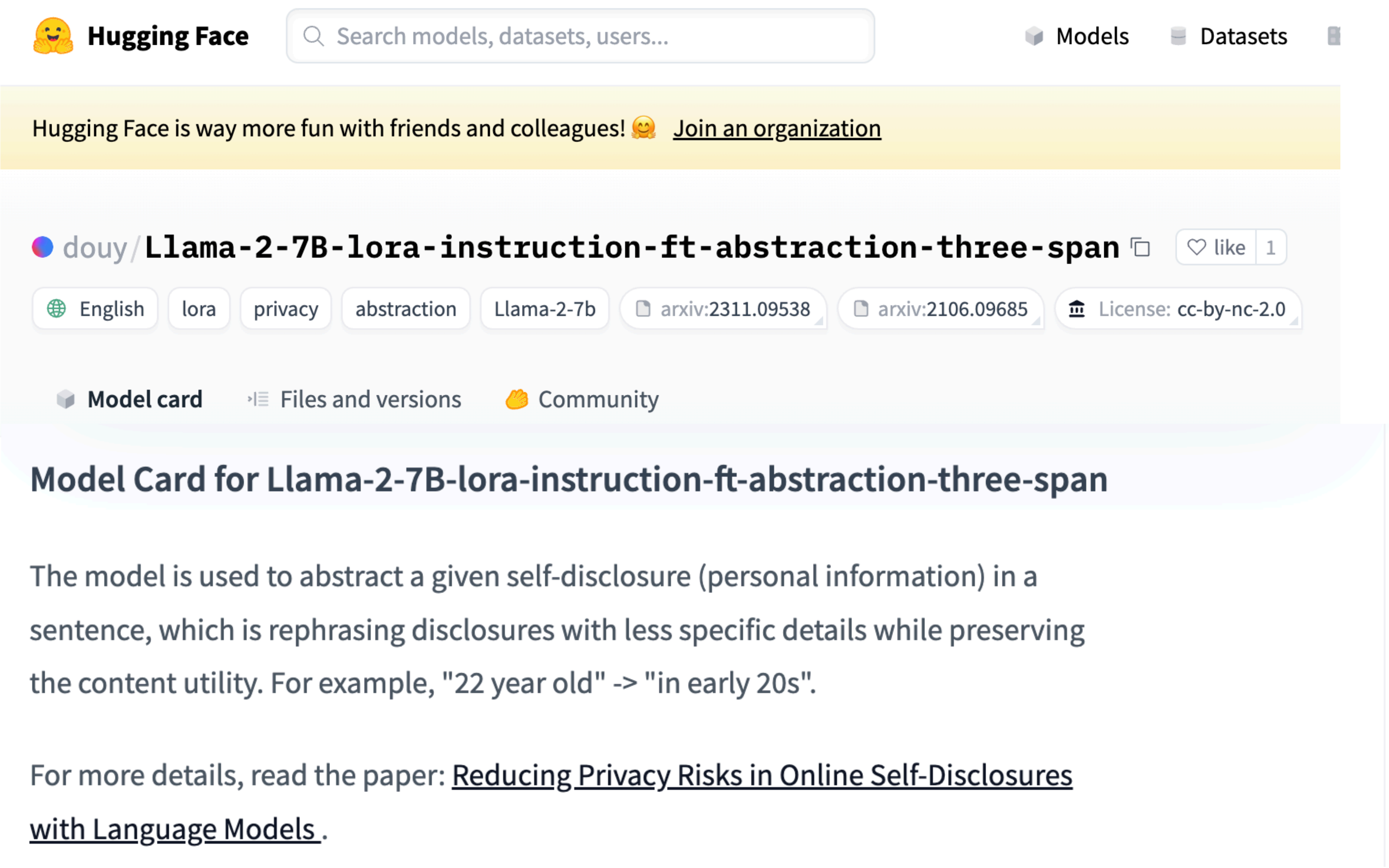
arXiv [cs.CL] 20 Feb 2024

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 F_1. 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., "Im

Im 16F I think I want to be a bi M
I am exploring my sexual identity
I have a desire to explore new options
I am attracted to the idea of exploring different gender identities
I live in New Mexico.
I live in the Southwest of the United States.
I reside in a state bordering Mexico.
I am currently living in a state known for its desert landscape.
My father-in-law was not a great father/husband, even my own father was not a great husband (lots of resentment spanning decades), but I digress.
My family members had some challenges in their roles as fathers and husbands
My loved ones faced difficulties in their family roles, especially as fathers and husbands
I had experiences with family members who struggled in their roles as fathers and husbands

Model on Huggingface



Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, Wei Xu. "Reducing Privacy Risks in Online Self-Disclosures with Language Models" (ACL 2024)

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

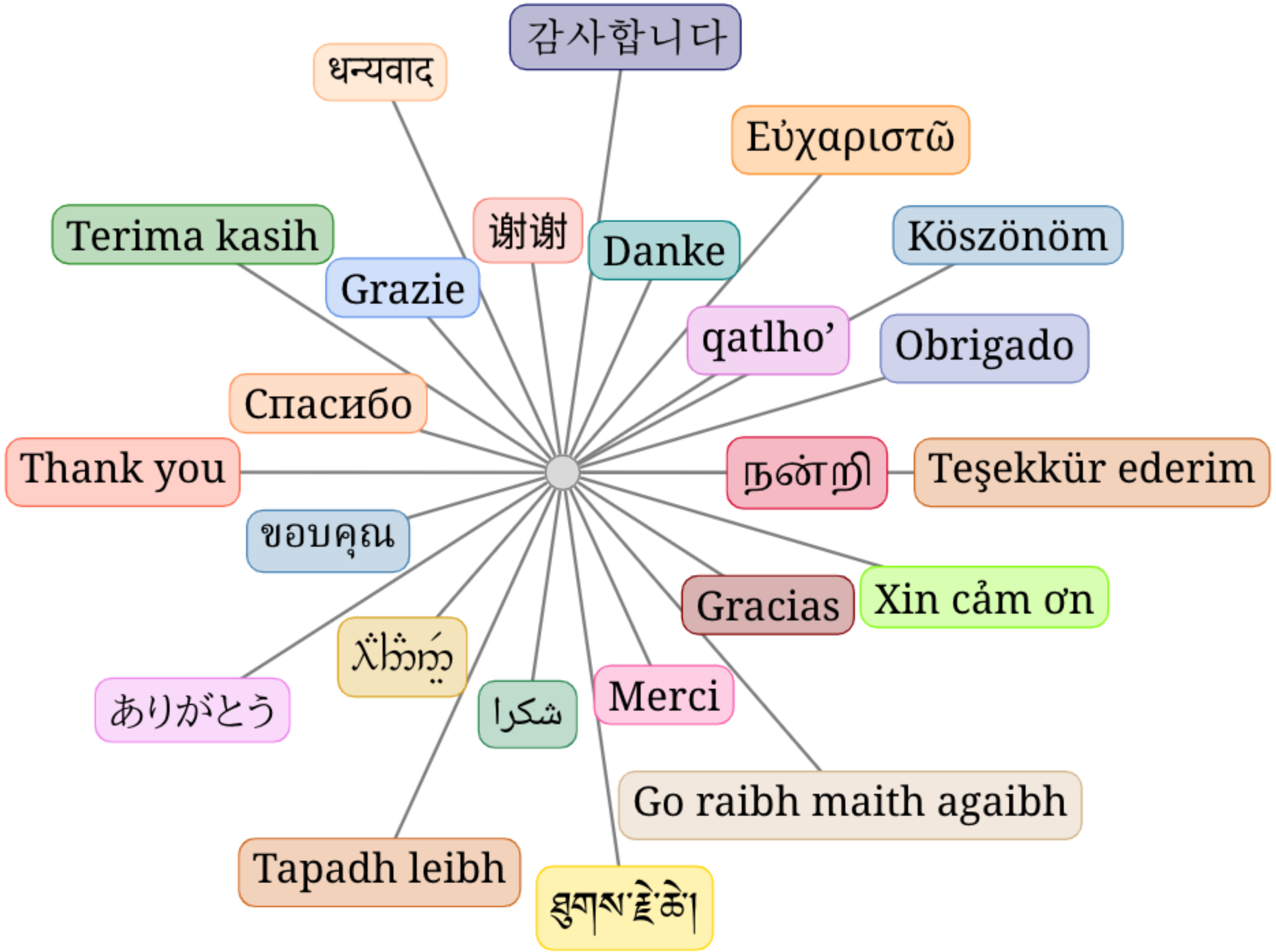
1. We need not only multilingual LLMs, but also multicultural LLMs.

2. 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.

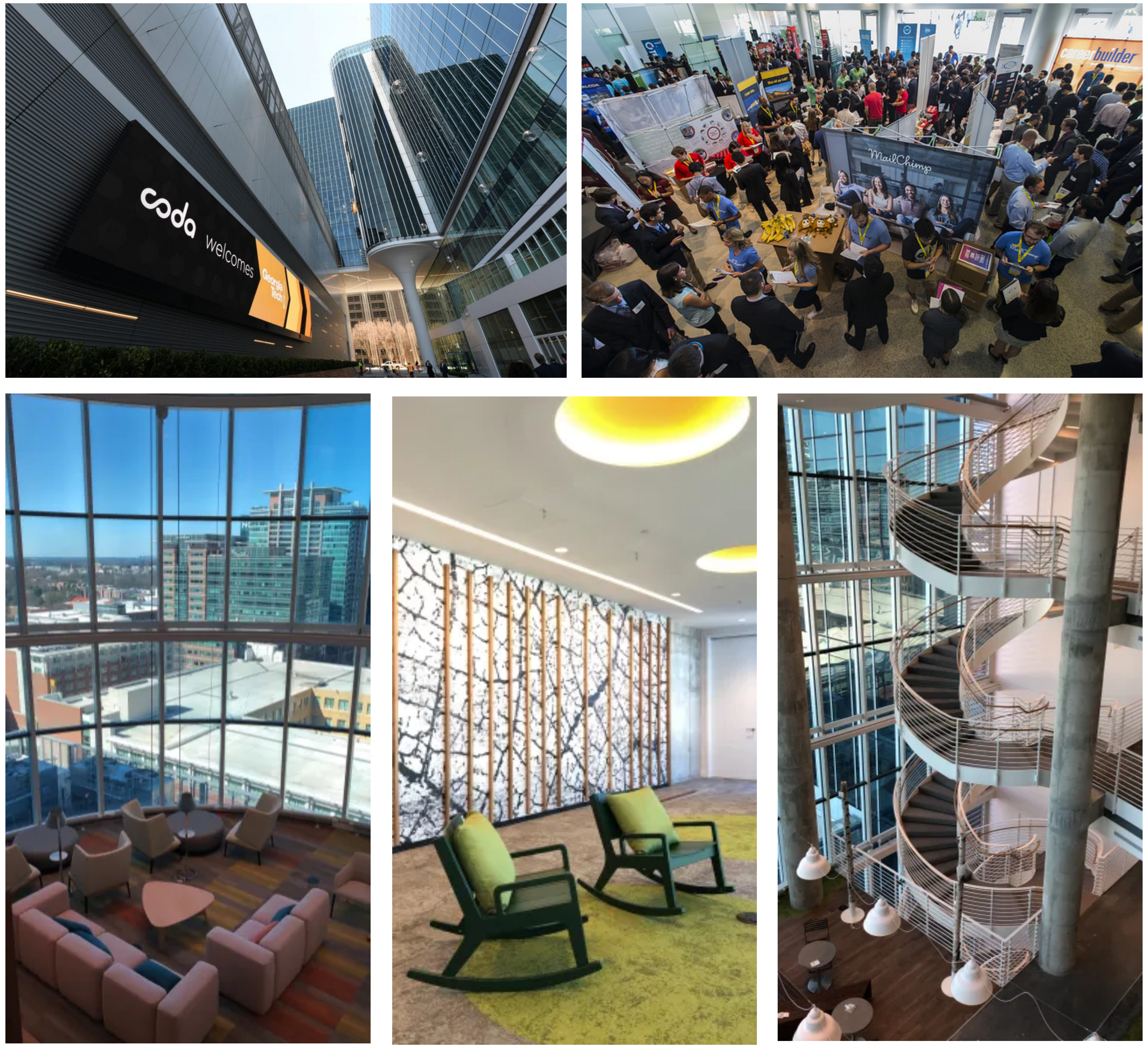
3. We will want to democratize the privacy protection, empowering individual end users to protect their own data.

Thank you!

https://cocoxu.github.io/



(image credit: Overleafl)



(image credit: Georgia Tech)

Georgia Tech Machine Learning

NSF

IARPA

DARPA