



# Improving Multilingual Capabilities and Error Detection 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

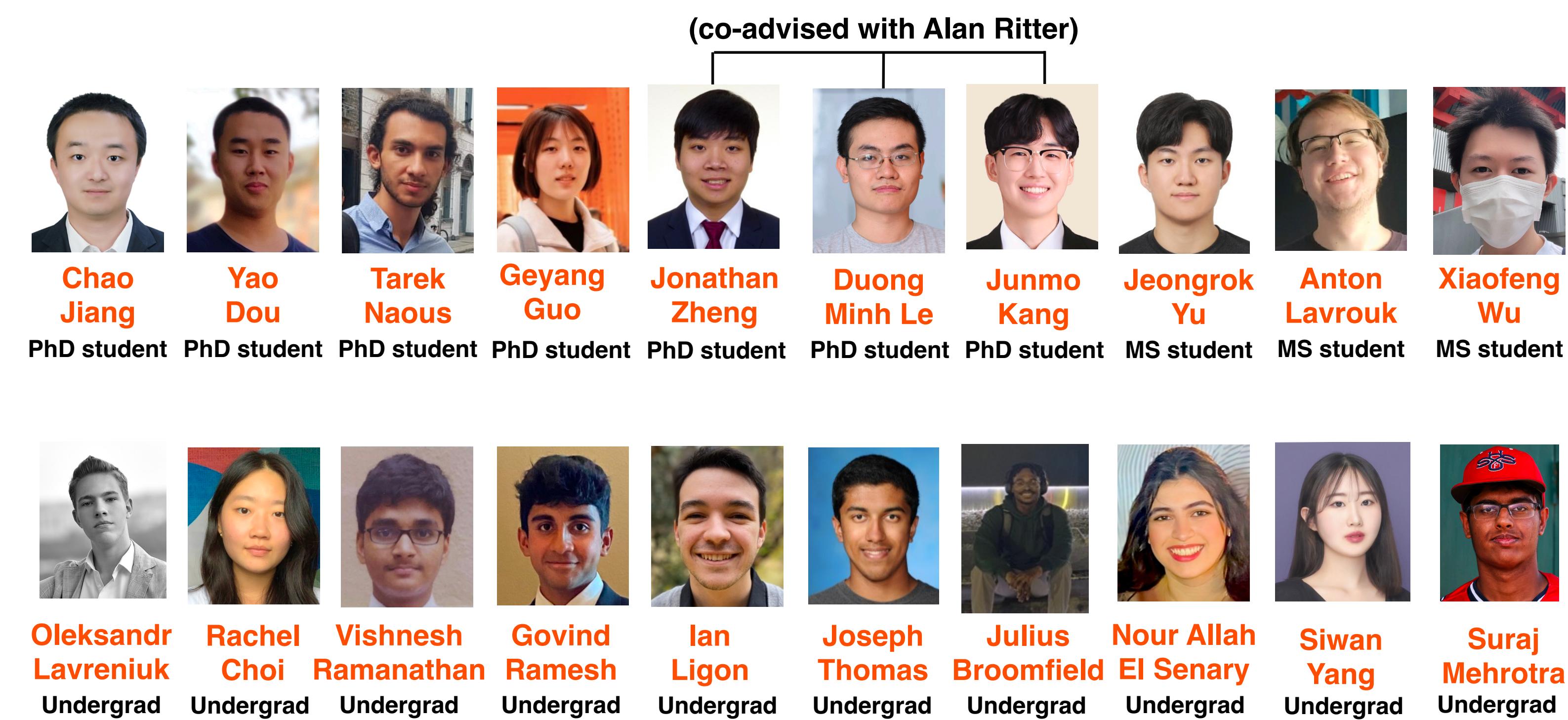
- evaluation of LLM-generated text
- reading/writing/voice assistant
- human-AI interactive system
- stylistics, personalization

## Language Models

- multi-/cross-lingual capability
- decoding algorithms
- cultural alignment
- privacy, safety
- reasoning

## NLP+X Interdisciplinary Research

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





# ACL 2024 Tutorial

## Automatic and Human-AI Interactive Text Generation

Sunday, August 11th 09:00 - 12:30, at Lotus 5-7 (Level 22)



Yao Dou<sup>\*1</sup>



Philippe Laban<sup>\*2</sup>



Claire Gardent<sup>3</sup>



Wei Xu<sup>1</sup>

<sup>1</sup>Georgia Institute of Technology, <sup>2</sup>Salesforce AI, <sup>3</sup>CNRS/LORIA

\*Equal Contribution

## Tutorial Schedule

The tutorial will be held on Aug 11th (time are based on UTC+7 = Bangkok local time):

Time	Section	Presenter
9:00 - 9:15	Introduction [ <a href="#">Slides</a> ]	Philippe
9:15 - 10:00	Section 1: Evaluation of LLM-generated Text [ <a href="#">Slides</a> ]	Yao
10:00 - 10:45	Section 2: Models - (Text Simplification and Text Rewriting) [ <a href="#">Slides</a> ]	Claire
10:45 - 11:00	Break	-
11:00 - 11:45	Section 3: Modeling Perspectives - (decoding, distillation, and diffusion LM) [ <a href="#">Slides</a> ]	Wei
11:45 - 12:30	Section 4: from method to usable system: Human-Centered NLP [ <a href="#">Slides</a> ]	Philippe

<https://acl2024-text-generation-tutorial.github.io/>

# Today's Talk —

## 1 - Multilingual LLMs & Decoding

CODEC



(Le et al., ICLR 2024)

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

## 2 - Evaluation of LLMs

CAMEL & Thresh



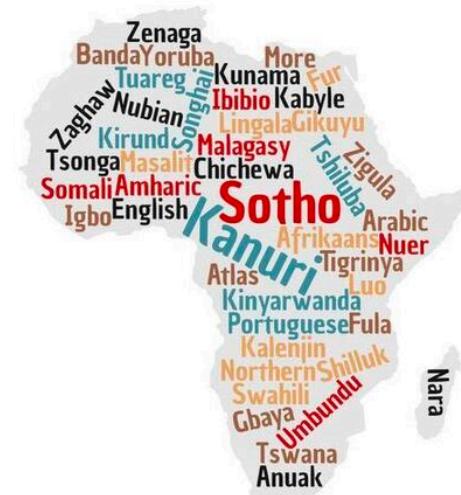
(Naous et al., EMNLP 2024 & Heineman et al., EMNLP 2023)

Evaluate what LLMs are good and bad at, for things that may or may not be very obvious.

# Today's Talk —

## 1 - Multilingual LLMs & Decoding

CODEC

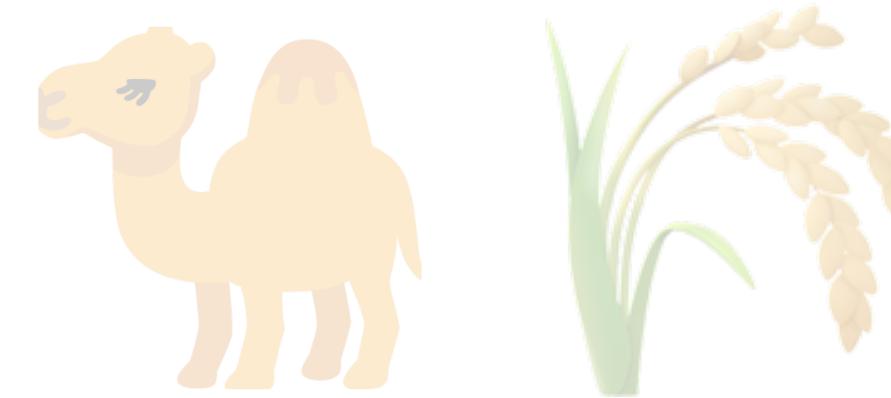


(Le et al., ICLR 2024)

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

## 2 - Evaluation of LLMs

CAMEL & Thresh



(Naous et al., EMNLP 2024 & Heineman et al., EMNLP 2023)

Evaluate what LLMs are good and bad at, for things that may or may not be very obvious.

# GPT-4 is surprisingly “good” at low resource languages

Masakhane NER 2.0



Lang.	GPT-4 <sup>†</sup>	FT <sub>En</sub>
Bambara	46.8	37.1
Ewe	75.5	75.3
Fon	19.4	49.6
Hausa	70.7	71.7
Igbo	51.7	59.3
Kinyarwanda	59.1	66.4
Luganda	73.7	75.3
Luo	<b>55.2</b>	35.8
Mossi	44.2	45.0
Chichewa	75.8	<b>79.5</b>
chiShona	66.8	35.2
Kiswahili	82.6	<b>87.7</b>
Setswana	62.0	64.8
Akan/Twi	52.9	50.1
Wolof	62.6	44.2
isiXhosa	69.5	24.0
Yoruba	<b>58.2</b>	36.0
isiZulu	60.2	43.9
AVG	60.4	54.5

# GPT-4 is surprisingly “good” at low resource languages

Masakhane NER 2.0



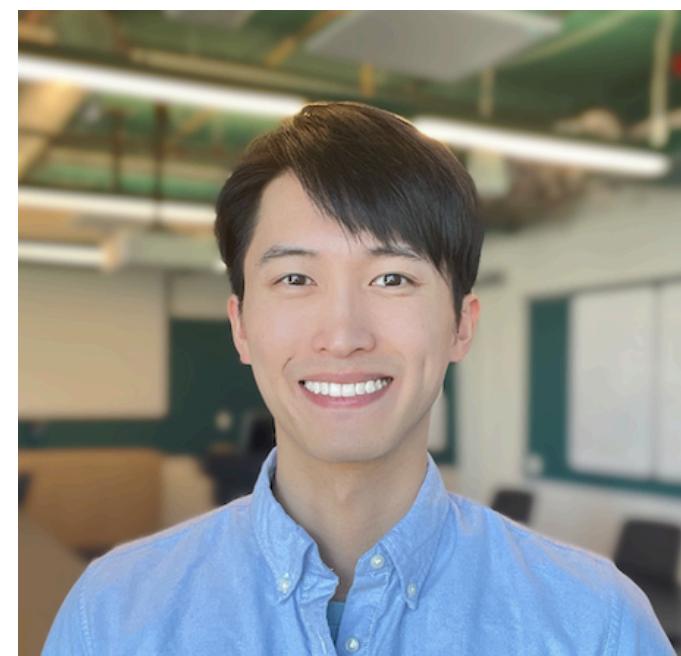
Lang.	GPT-4 <sup>†</sup>	FT <sub>En</sub>	Translate-train			Translate-test	
			Awes-align	EasyProject	CODEC ( $\Delta_{FT}$ )	Awes-align	CODEC ( $\Delta_{FT}$ )
Bambara	46.8	37.1	45.0	45.8	45.8 (+8.7)	50.0	<b>55.6</b> (+18.5)
Ewe	75.5	75.3	78.3	78.5	<b>79.1</b> (+3.8)	72.5	<b>79.1</b> (+3.8)
Fon	19.4	49.6	59.3	61.4	<b>65.5</b> (+15.9)	62.8	61.4 (+11.8)
Hausa	70.7	71.7	72.7	72.2	72.4 (+0.7)	70.0	<b>73.7</b> (+2.0)
Igbo	51.7	59.3	63.5	65.6	70.9 (+11.6)	<b>77.2</b>	72.8 (+13.5)
Kinyarwanda	59.1	66.4	63.2	71.0	71.2 (+4.8)	64.9	<b>78.0</b> (+11.6)
Luganda	73.7	75.3	77.7	76.7	77.2 (+1.9)	<b>82.4</b>	82.3 (+7.0)
Luo	<b>55.2</b>	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	<b>55.6</b> (+10.6)	48.4	50.4 (+5.4)
Chichewa	75.8	<b>79.5</b>	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	<b>78.4</b> (+43.2)
Kiswahili	82.6	<b>87.7</b>	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)	<b>81.4</b>	80.3 (+15.5)
Akan/Twi	52.9	50.1	62.7	65.3	64.6 (+14.5)	72.6	<b>73.5</b> (+23.4)
Wolof	62.6	44.2	54.5	58.9	63.1 (+18.9)	58.1	<b>67.2</b> (+23.0)
isiXhosa	69.5	24.0	61.7	<b>71.1</b>	70.4 (+46.4)	52.7	69.2 (+45.2)
Yoruba	<b>58.2</b>	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	<b>74.8</b> (+30.9)	64.1	<b>76.9</b> (+33.0)
AVG	60.4	54.5	63.6	64.9	67.1 (+12.7)	65.8	<b>70.4</b> (+16.0)

Label Projection  
methods

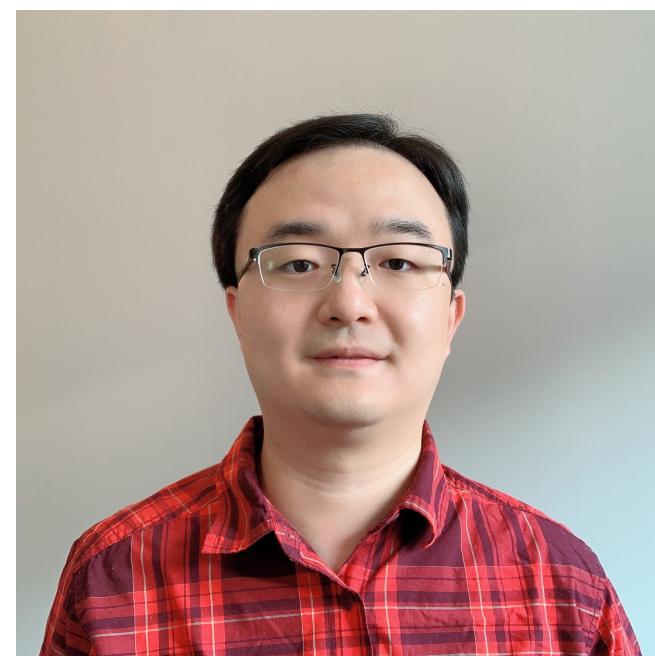
[1] CODEC: “Constrained Decoding for Cross-lingual Label Projection” Duong Minh Le, Yang Chen, Alan Ritter, Wei Xu (ICLR 2024)

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

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



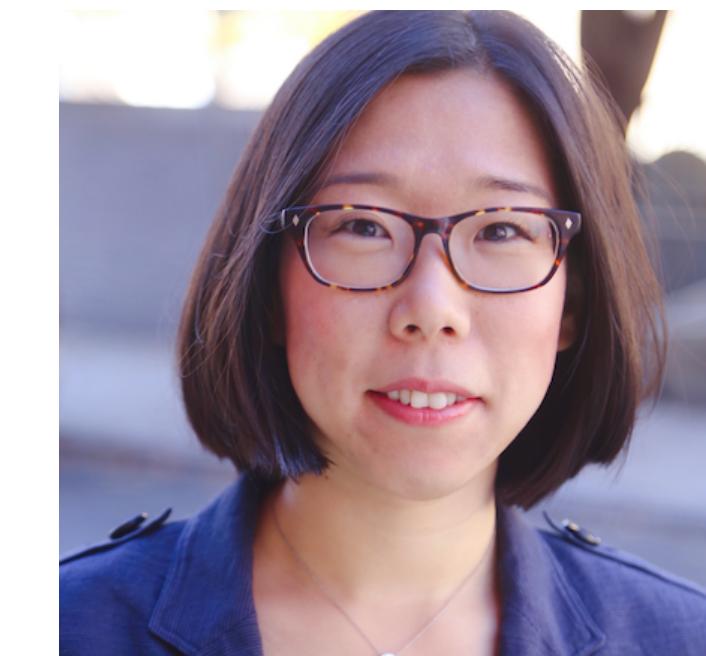
Yang Chen



Chao Jiang



Alan Ritter

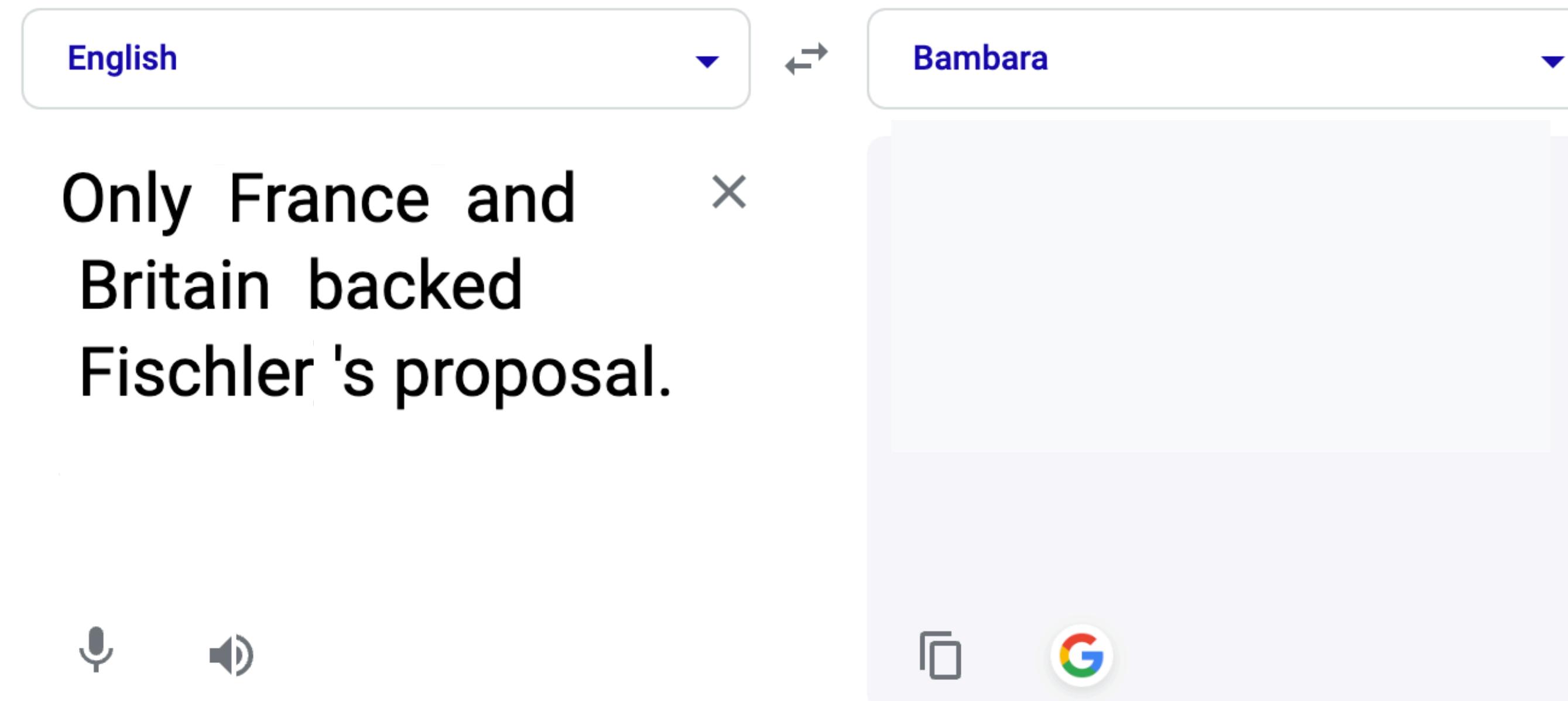


Wei Xu

A systematic study of marker-based  
approach for label projection

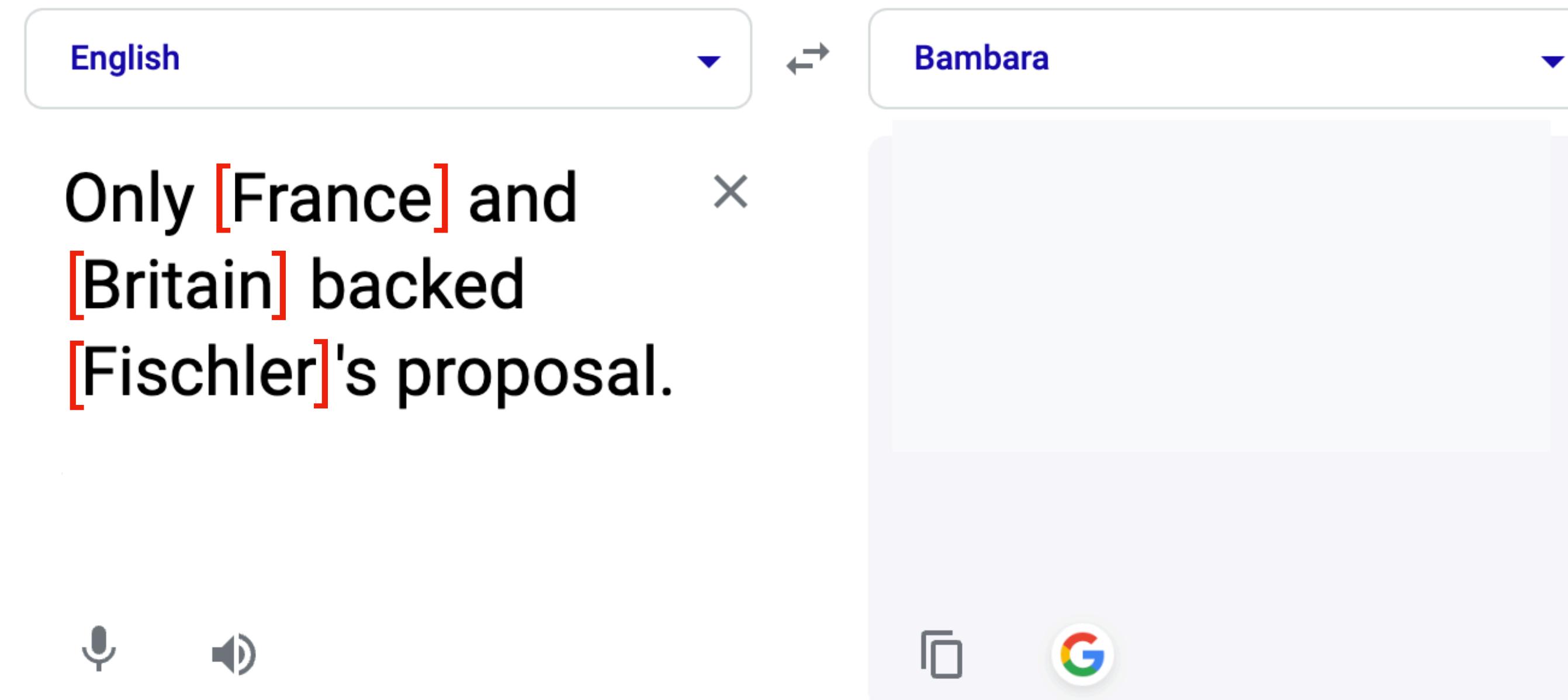
# Marker-based Approach

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. Between them is a double-headed arrow icon. Below the dropdowns is a text input field containing the sentence: 'Only [France] and [Britain] backed [Fischler]'s proposal.' The words 'France', 'Britain', and 'Fischler' are highlighted with red brackets. To the right of this input field is a large gray button with the translated text: '[France] ni [Britagne] dɔrɔn de ye [Fischler] ka lajini dɛmɛ.' At the bottom of the screen, there are several icons: a microphone icon, a speaker icon, a refresh/circular arrow icon, 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, Refresh icon, Google icon

# 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

# 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

# EasyProject - Easy Marker-based Projection

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

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

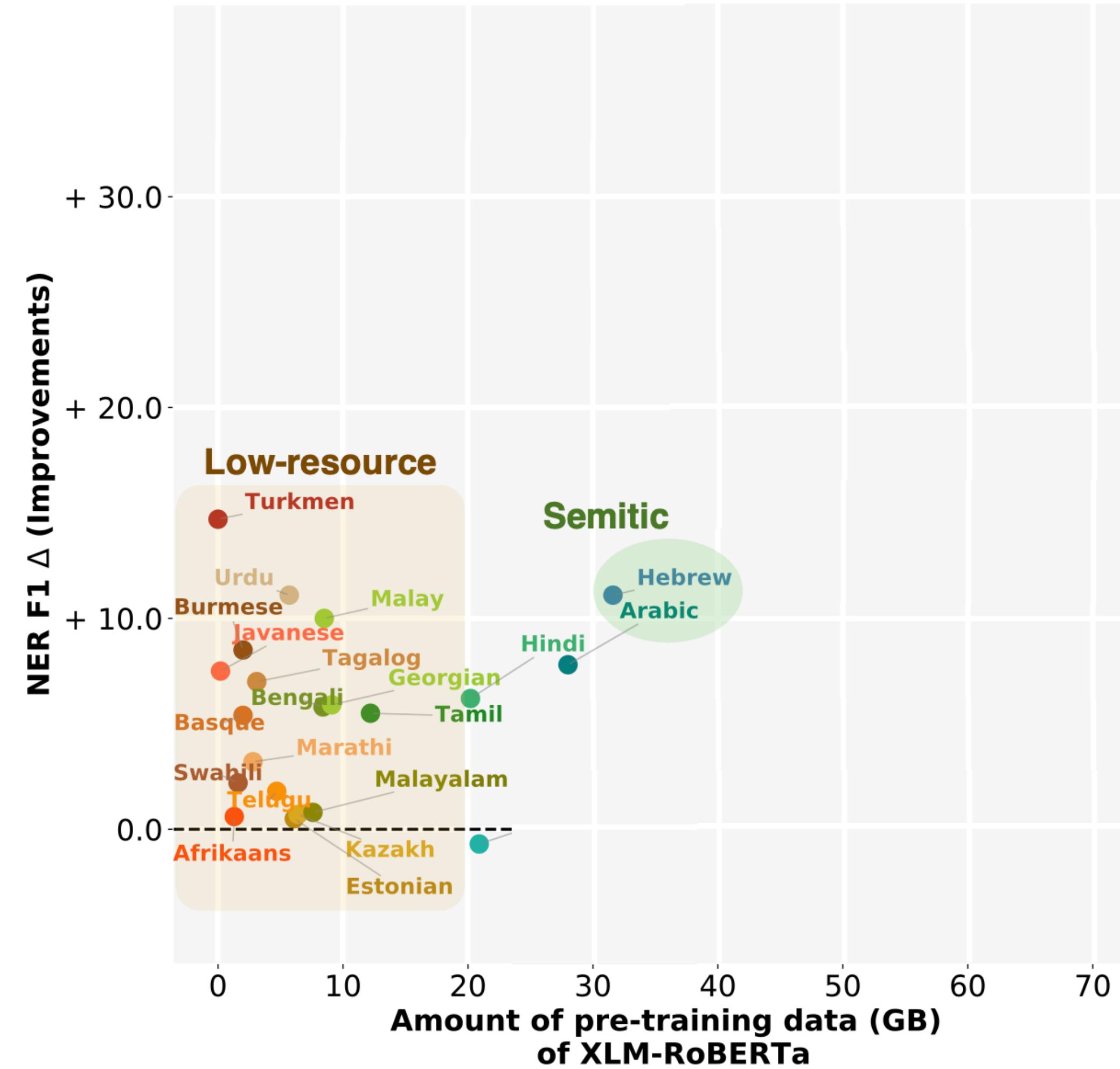
works the best



- 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

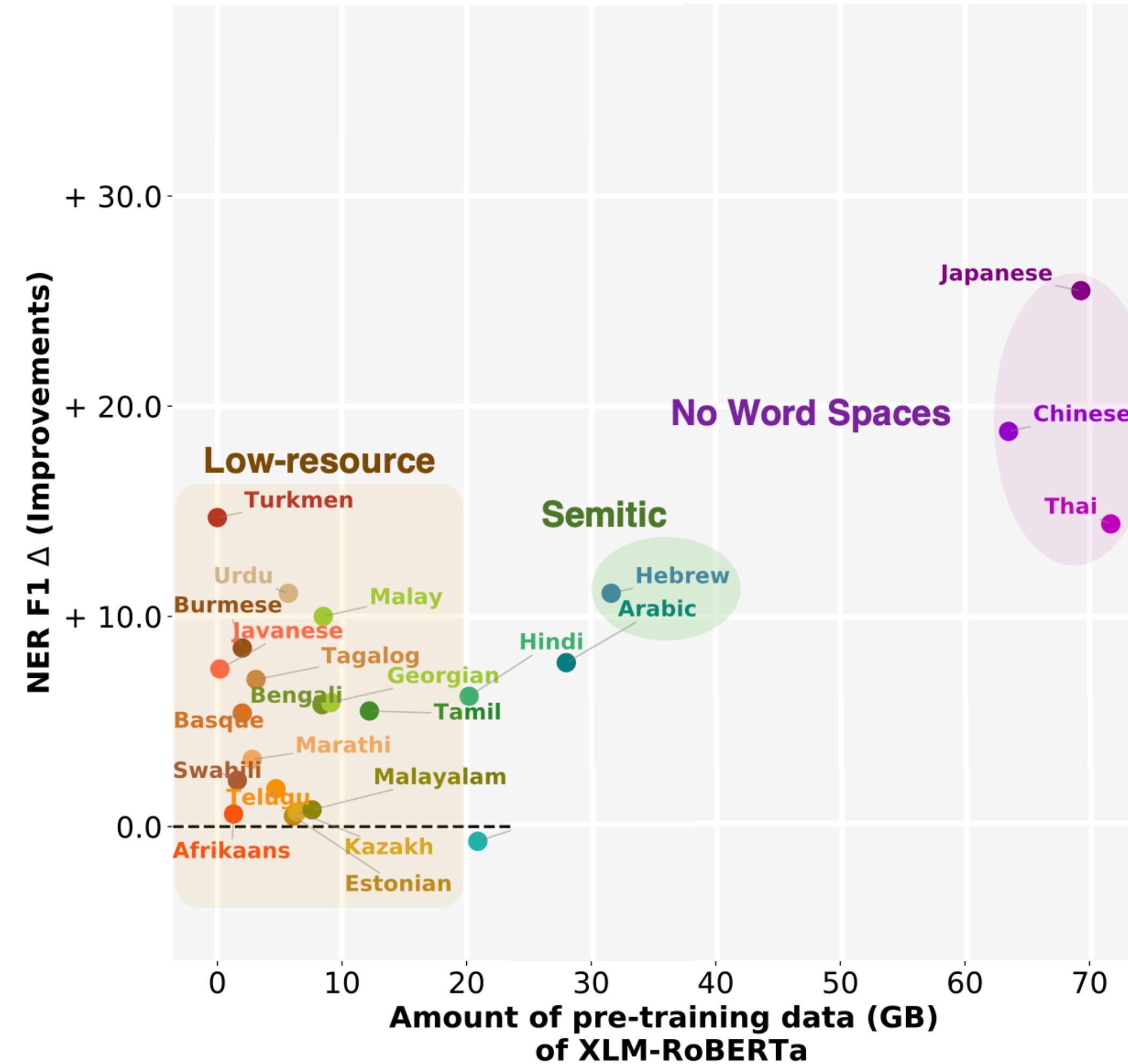
# EasyProject - Easy Marker-based Projection

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



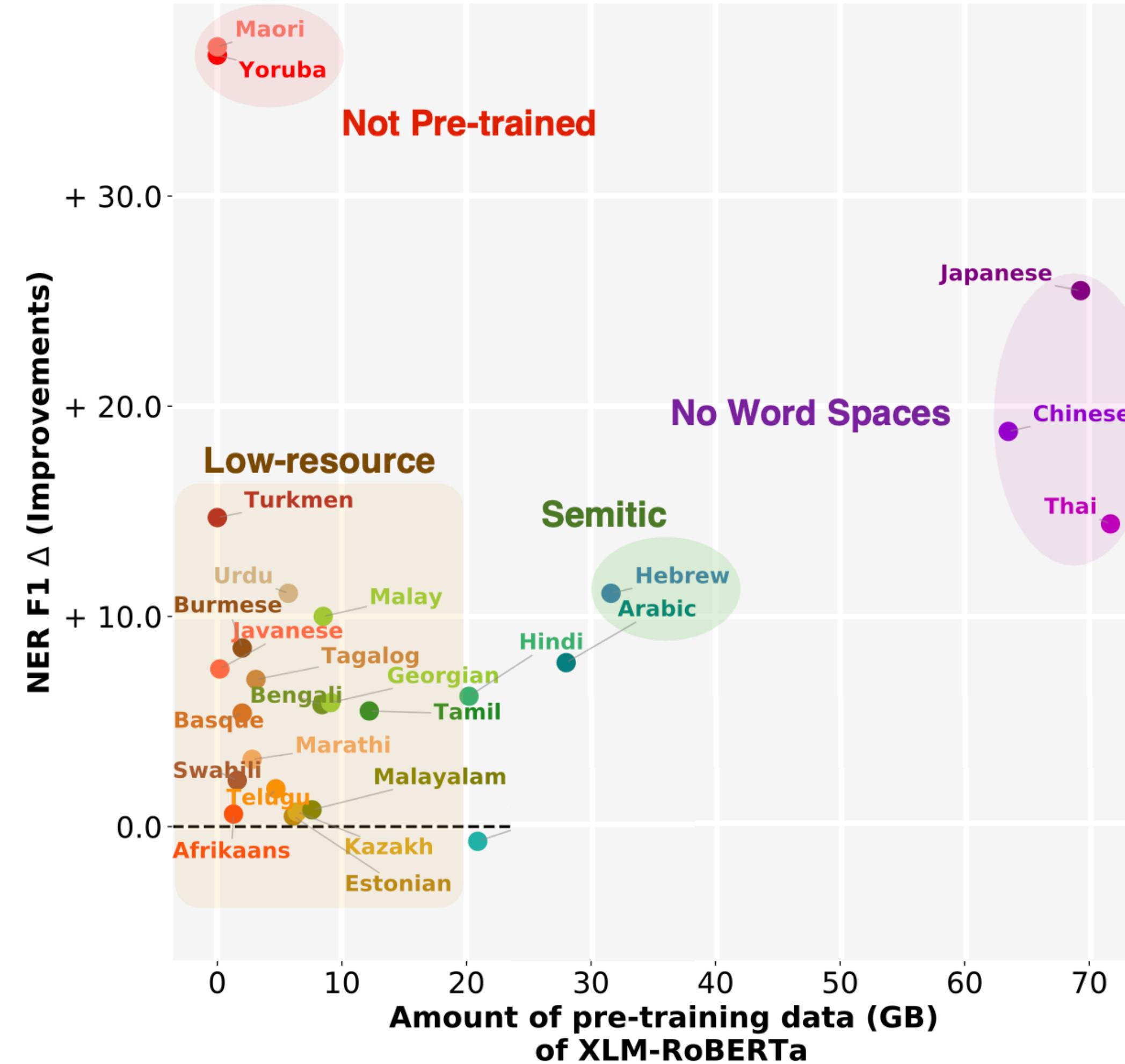
# EasyProject - Easy Marker-based Projection

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



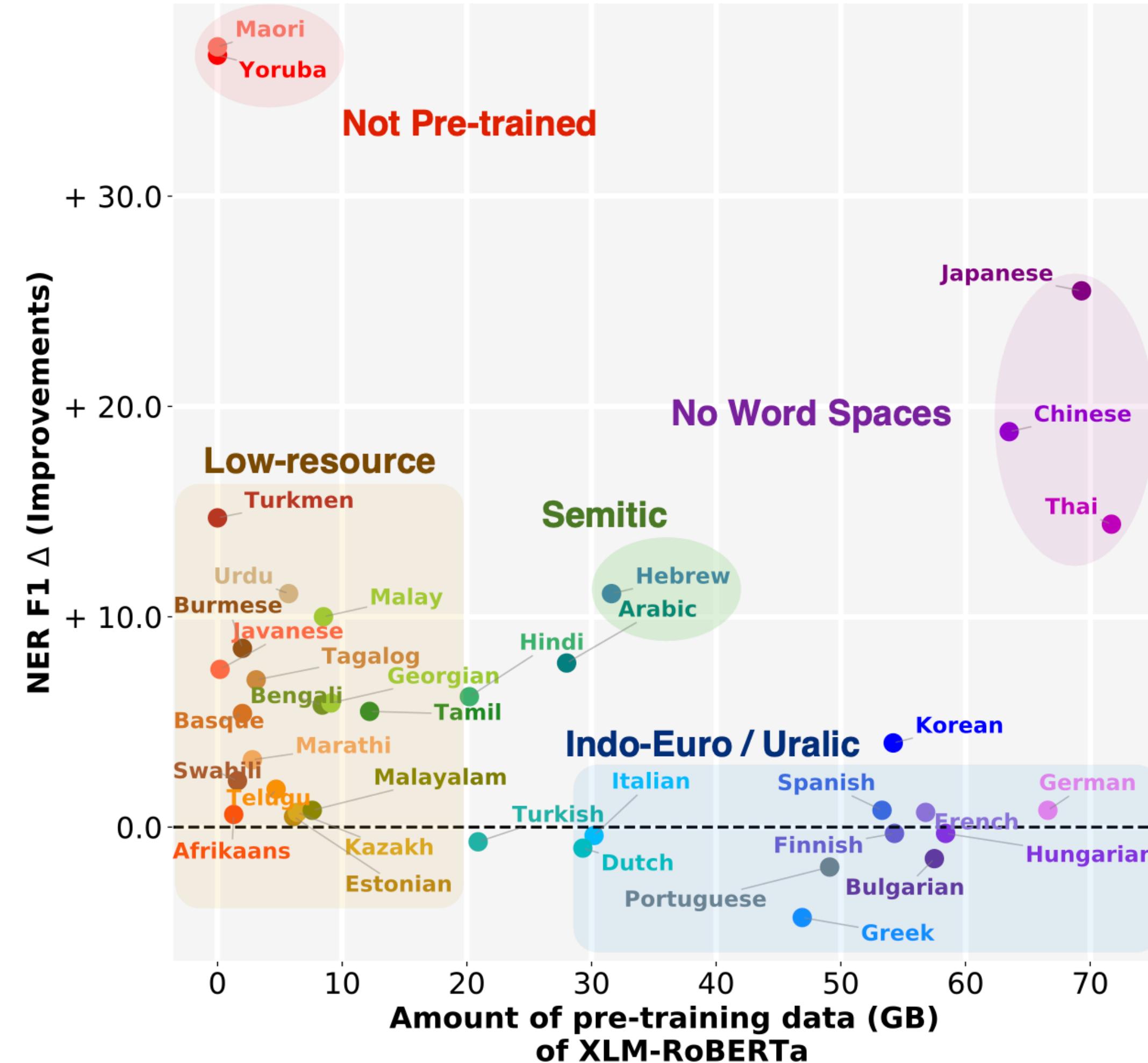
# EasyProject - Easy Marker-based Projection

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

English ▾ ↔ Bambara ▾

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

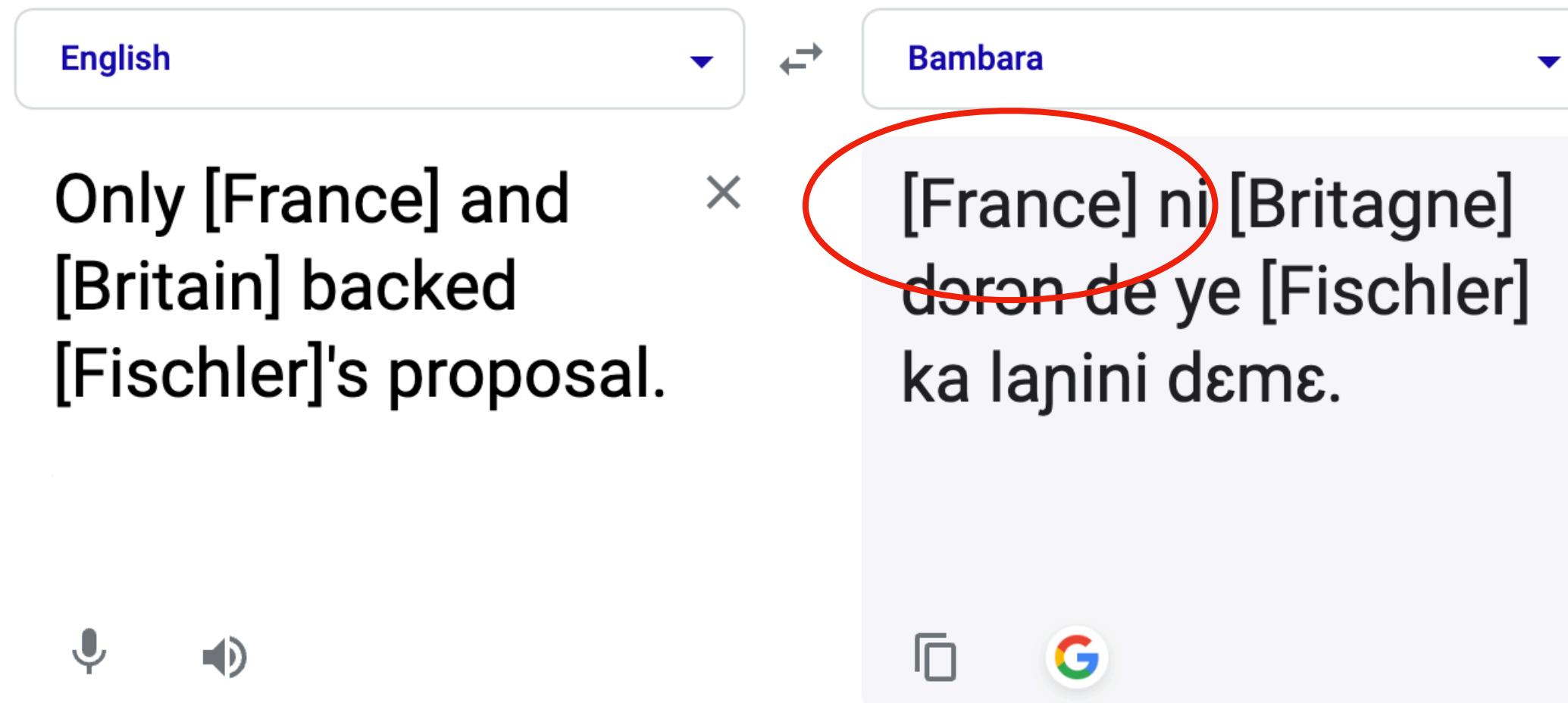
[France] ni [Britagne]  
dɔrɔn de ye [Fischler]  
ka lañini dɛmɛ.

Microphone icon, Speaker icon, Google search icon

# 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 machine translation interface with two dropdown menus for "English" and "Bambara". Below the English input, the text "Only [France] and [Britain] backed [Fischler]'s proposal." is displayed. In the Bambara output, the words "[France]" and "[Britain]" are circled in red. The full output is: "Only [France] ni [Britain] backed [Fischler]'s proposal. Faransi ni Angleterei dɔrɔn de ye Fischler ka lajini dɛmɛ .". There are also small icons for microphone, speaker, and Google.

Only need a MT system  
&  
work surprisingly well !

But, degraded  
MT quality  
due to injected markers

## word alignment-based approach

The screenshot shows a machine translation interface with two dropdown menus for "English" and "Bambara". Below the English input, the text "Only France and Britain backed Fischler's proposal ." is displayed. In the Bambara output, the words "Faransi" and "Angleterei" are underlined and aligned with "Only France" and "Britain" respectively. The full output is: "Only France and Britain backed Fischler's proposal . Faransi ni Angleterei dɔrɔn de ye Fischler ka lajini dɛmɛ .". A red box highlights the word alignment diagram above the output. The diagram shows arrows pointing from "Only France" to "Faransi", "and" to "ni", "Britain" to "Angleterei", "backed" to "dɔrɔn", "Fischler's" to "Fischler", and "proposal" to "lajini". The labels "LOC", "LOC", and "PER" are placed above the corresponding words in the English sentence.

# 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 source and target languages: "English" and "Bambara". The source sentence is "Only [France] and [Britain] backed [Fischler]'s proposal." The target sentence is "[France] ni [Britagne] dɔrɔn de ye [Fischler] ka lajini dɛmɛ." A red oval highlights the markers "[France]" and "[Britagne]" in the target sentence. Below the interface are two buttons: a microphone icon and a speaker icon.

Only need a MT system  
&  
work surprisingly well !

But, degraded  
MT quality  
due to injected markers

## word alignment-based approach

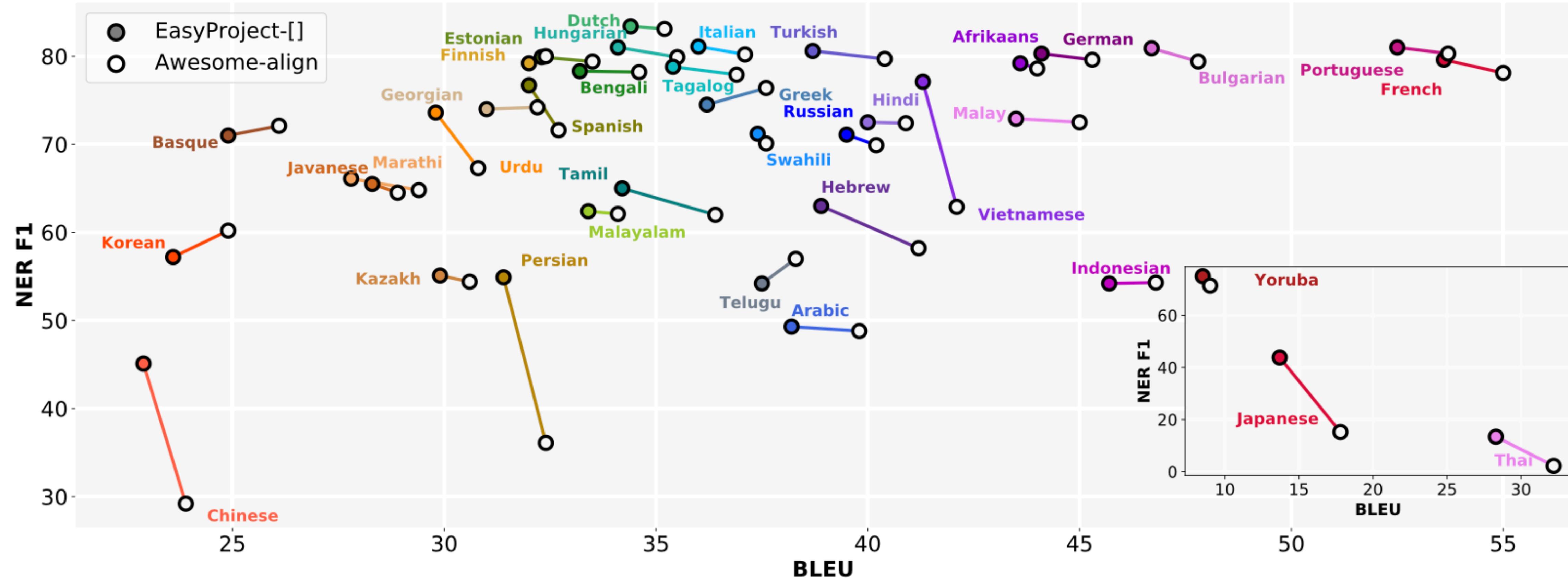
The screenshot shows a neural machine translation interface with two dropdown menus for source and target languages: "English" and "Bambara". The source sentence is "Only France and Britain backed Fischler's proposal." The target sentence is "Faransi ni Angleteri dɔrɔn de ye Fischler ka lajini dɛmɛ." Above the sentences, part-of-speech tags are shown: "LOC Only France and LOC Britain backed PER Fischler's proposal." Below the sentences, a red box highlights the words "France", "Britain", "Fischler", and "proposal" in both the source and target sentences, with arrows indicating their alignment. Below the interface are two buttons: a microphone icon and a speaker icon.

normally  
better MT quality

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

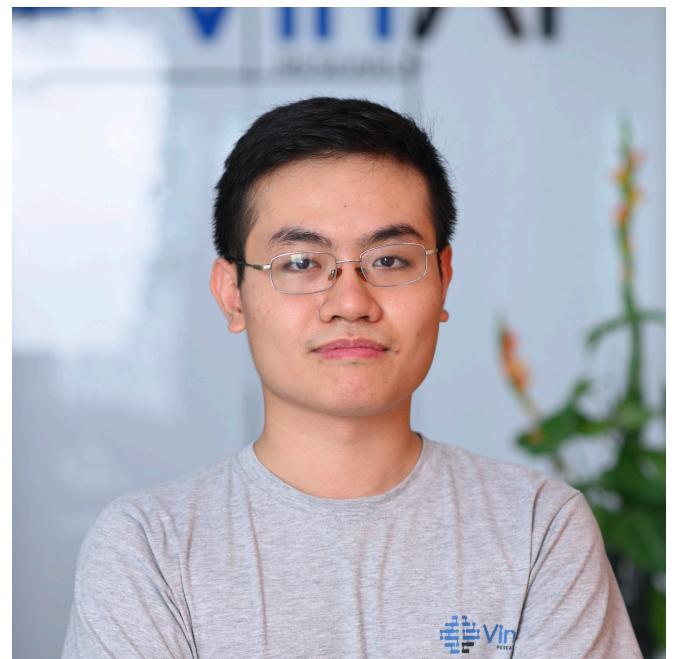
# EasyProject - Easy Marker-based Projection

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

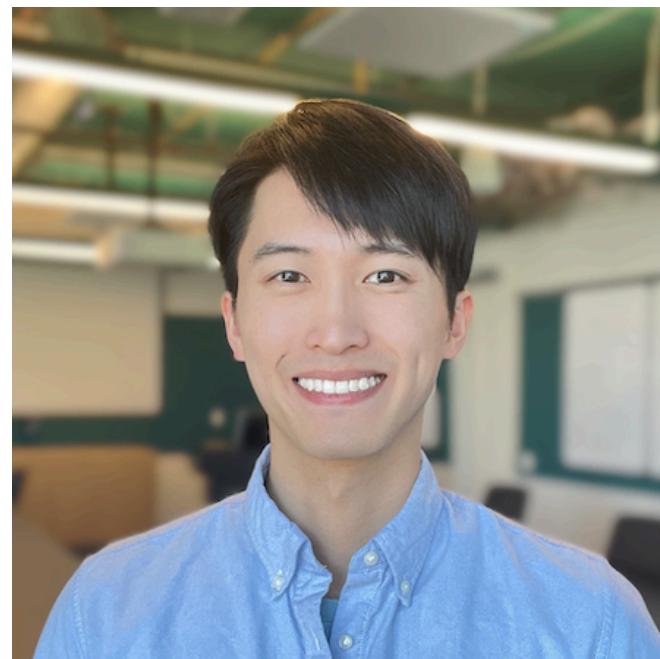


**Can we do marker-based approach without  
sacrificing the translation quality?**

# 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.

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 inputted: "Only France and Britain backed Fischler's proposal." A red "X" icon is positioned next to the input field. To the right, the Bambara translation is displayed: "Faransi ni Angleteri dɔrɔn de ye Fischler ka lanini dɛmɛ." At the bottom of the interface, there are three icons: a microphone for voice input, a speaker for audio output, and a refresh symbol.

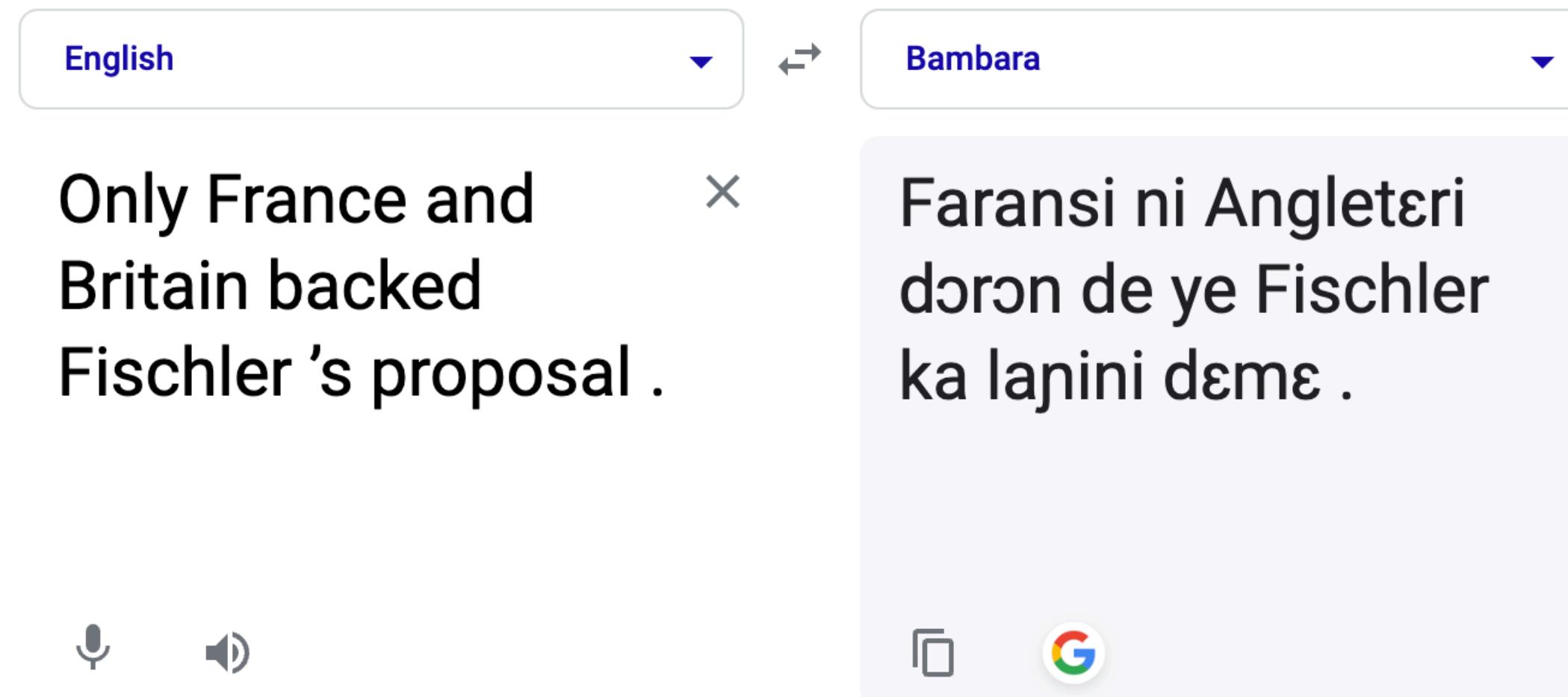
Step 2. Run translation model for a 2nd time to insert markers as a constrained decoding problem.

**Input sentence:**

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

# Key Idea

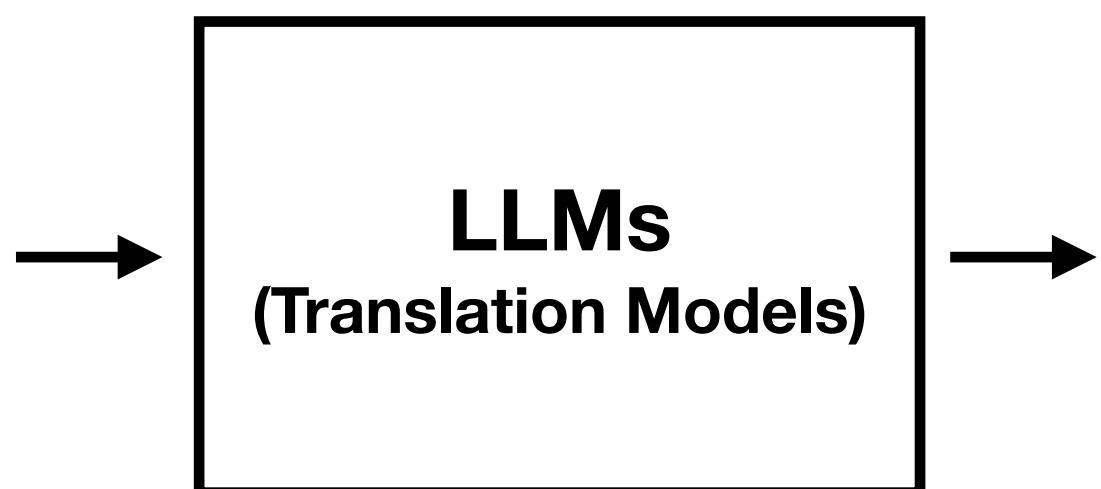
Step 1. Translate the original sentence as usual without markers.



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

**Input sentence:**

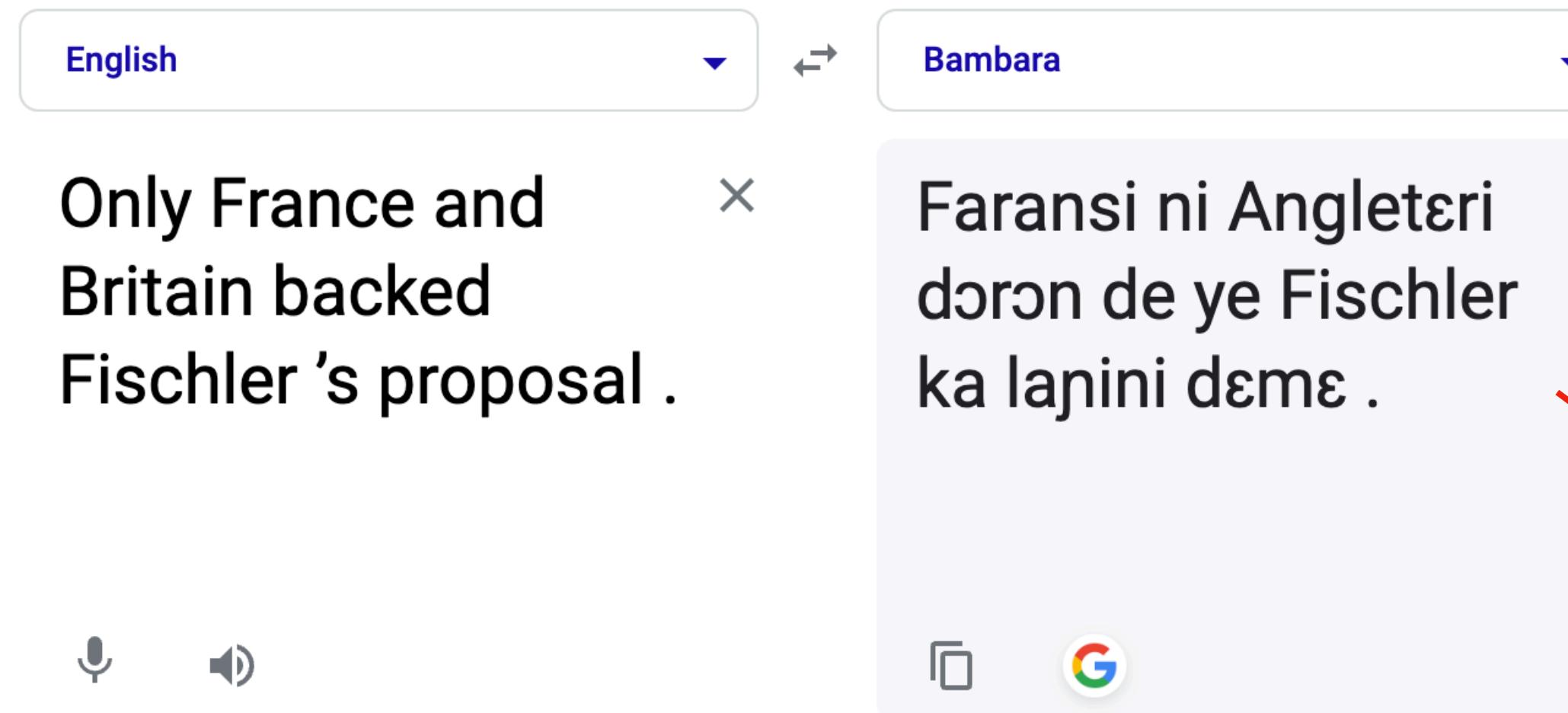
Only [France] and [Britain] backed [Fischler]'s proposal.



**Translated Output:**

# Key Idea

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

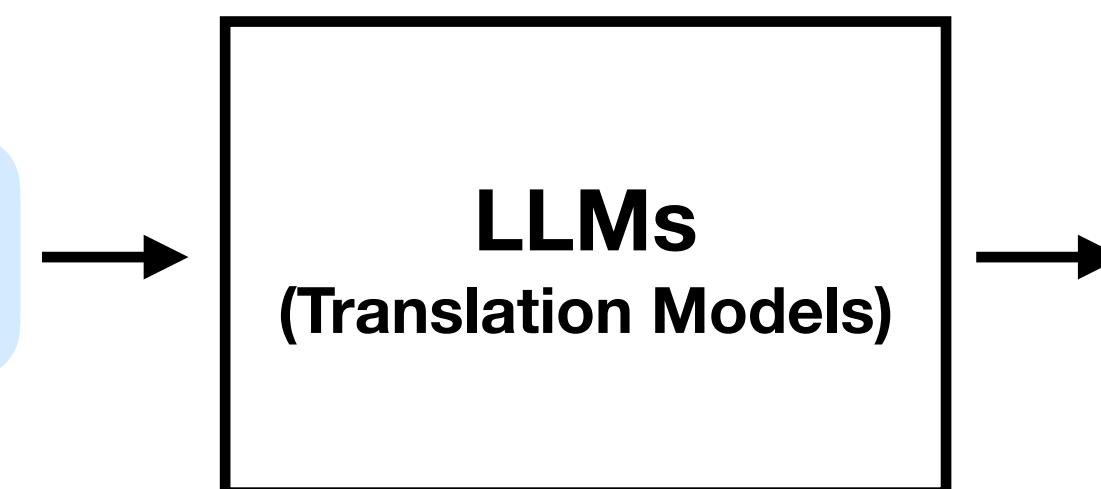


Impose two constraints:  
(1) keeping the same translation  
(2) having the correct number of [ ] s

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

**Input sentence:**

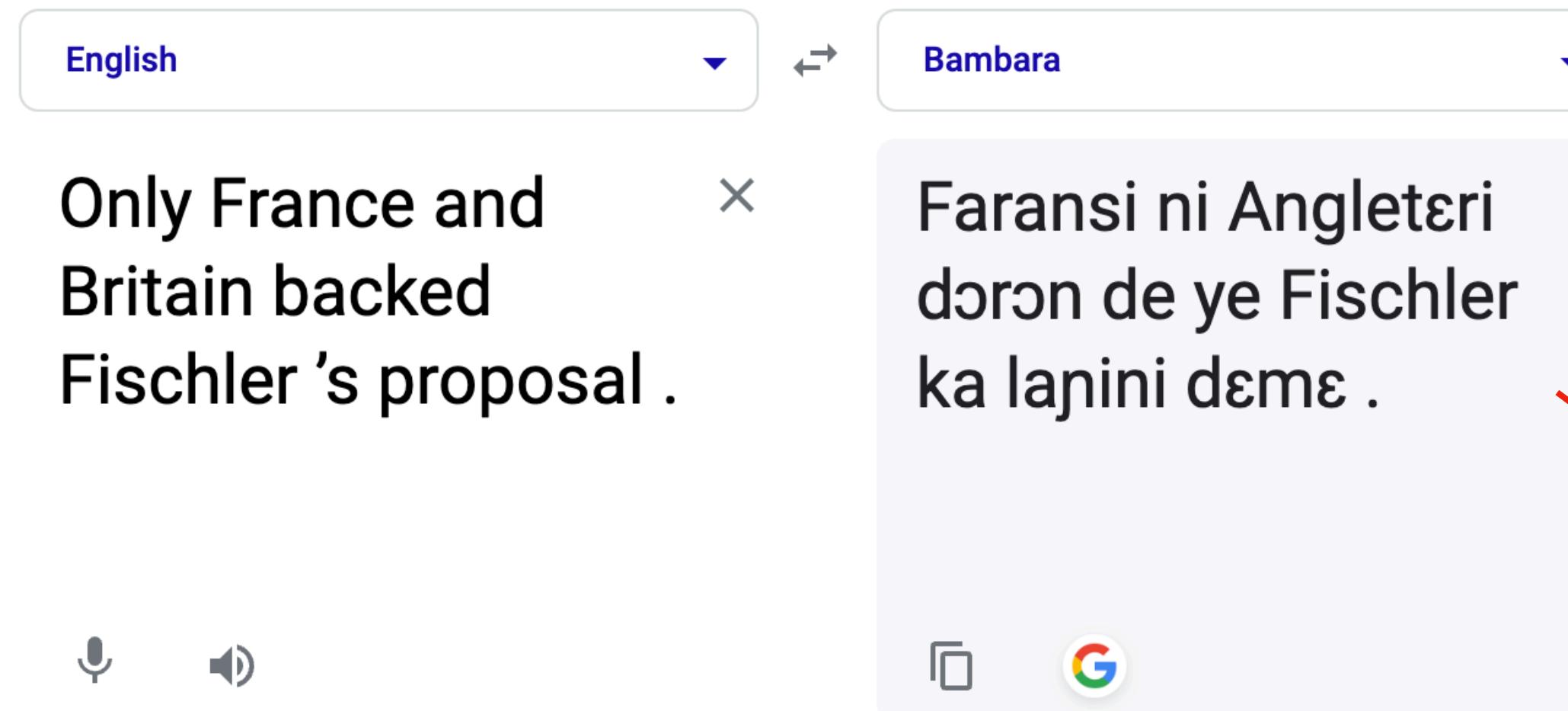
Only [France] and [Britain] backed [Fischler]'s proposal.



**Translated Output:**

# 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.

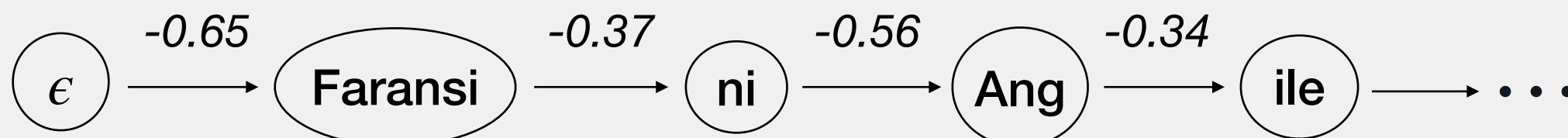
**Input:**

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

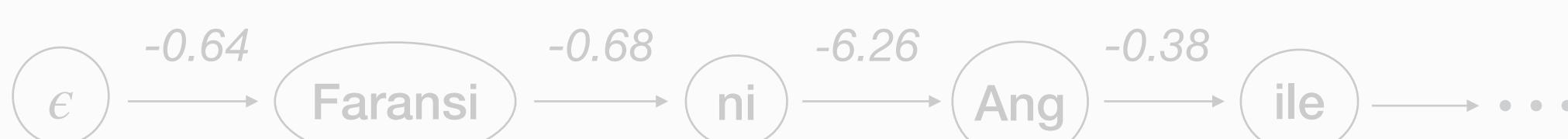
$x^{mark} = \text{"Only France and [ Britain ] backed Fischler 's proposal ."}$

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

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



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



# An Efficient Constrained Decoding Algorithm

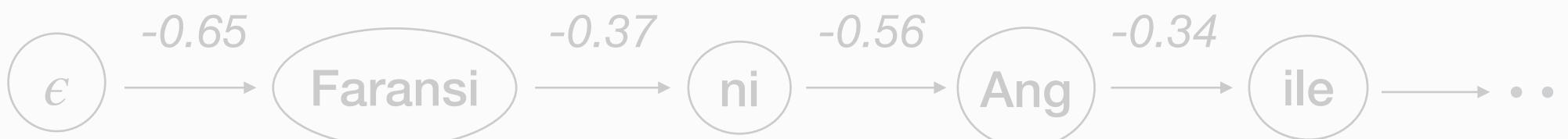
(1) Prune opening marker positions based on the contrastive log-likelihood difference.

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

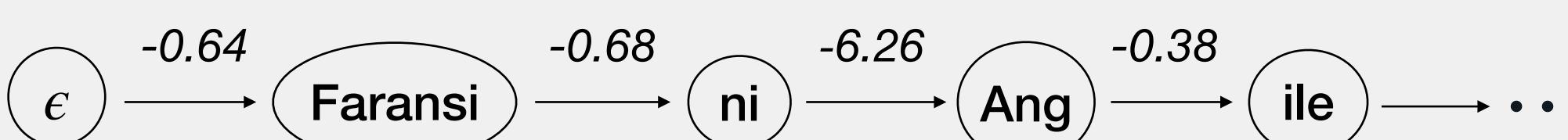
$x^{mark} = \text{"Only France and [ Britain ] backed Fischler 's proposal ."}$

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

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



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

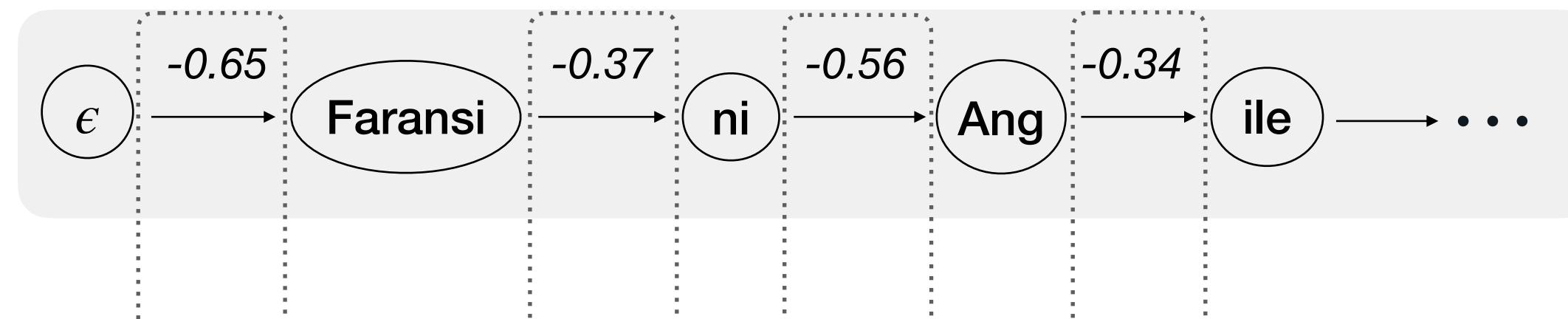


# An Efficient Constrained Decoding Algorithm

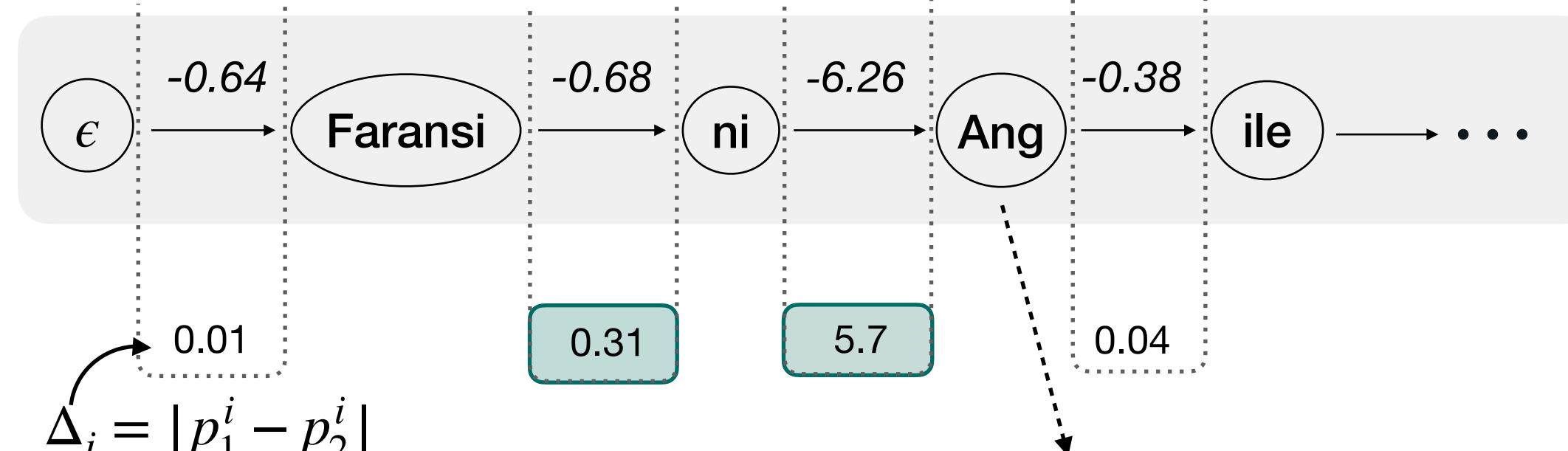
(1) Prune opening marker positions based on the contrastive log-likelihood difference.

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

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



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



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

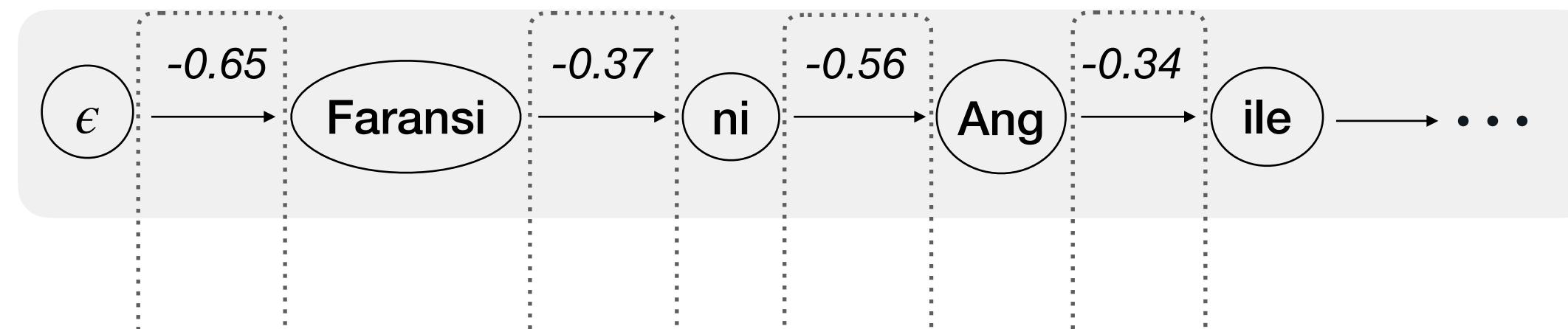
*This position should be '[', thus the transition probability is extremely low*

# An Efficient Constrained Decoding Algorithm

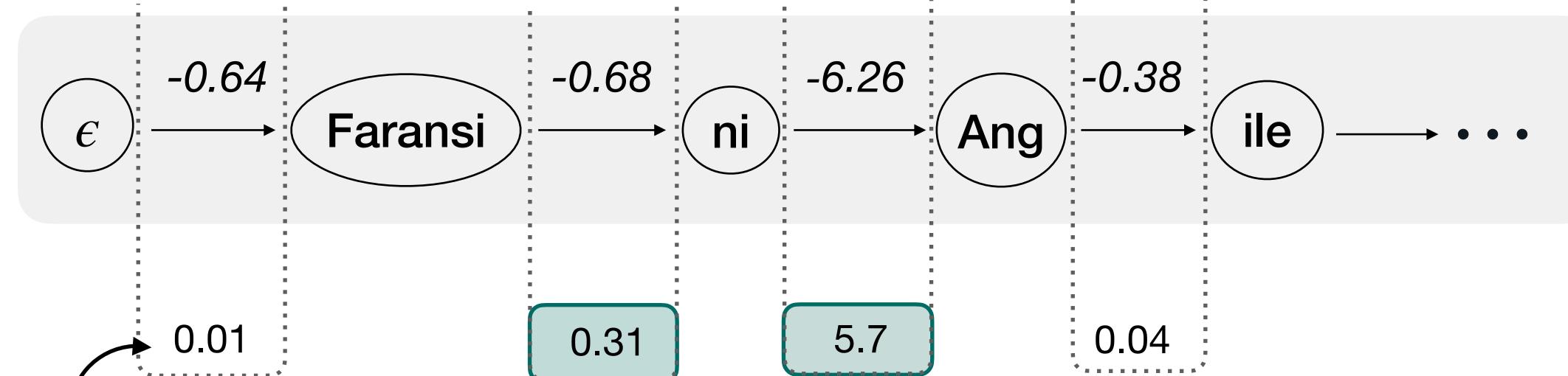
(1) Prune opening marker positions based on the contrastive log-likelihood difference.

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

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



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



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

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  $L_d^k = \log P(y_{1:d}^k | x^{mark})$  .  
 $d = \min (\max (j + \delta, q), |y^k|)$

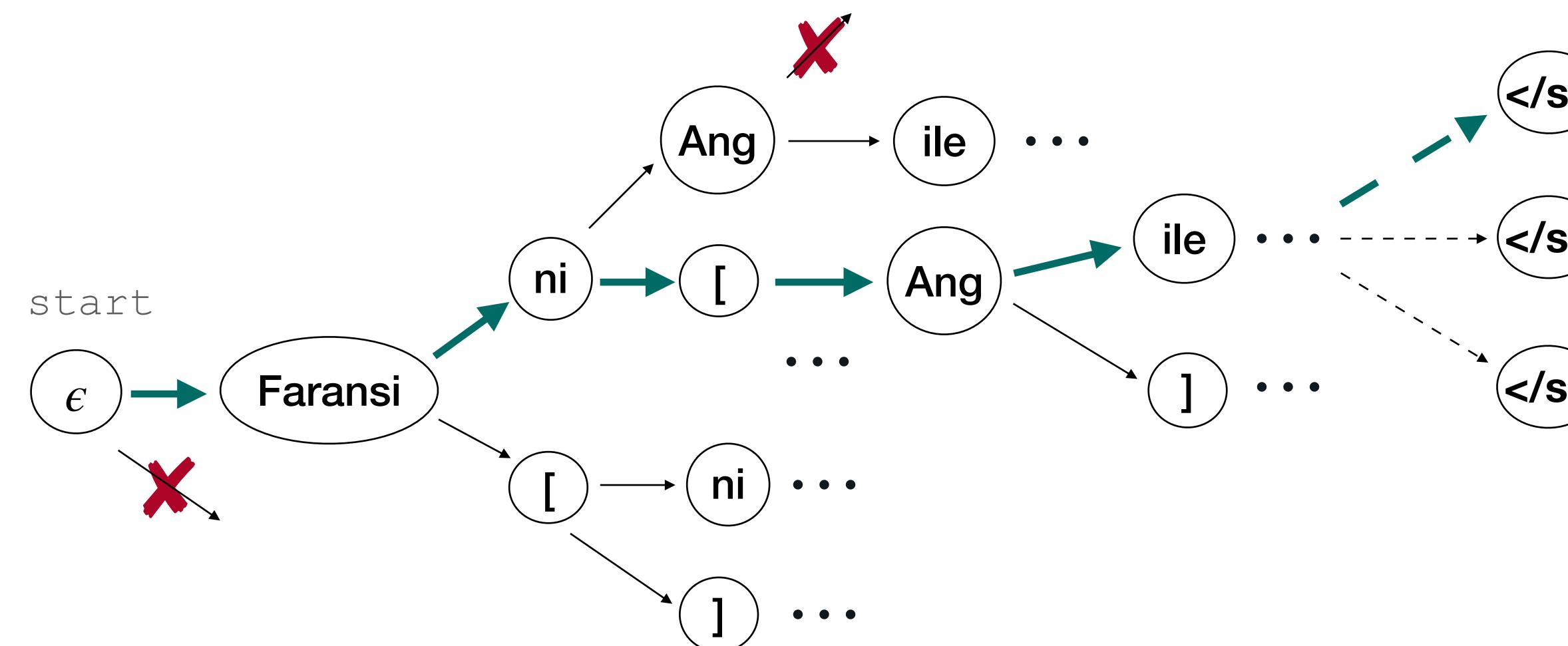
# An Efficient Constrained Decoding Algorithm

(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*

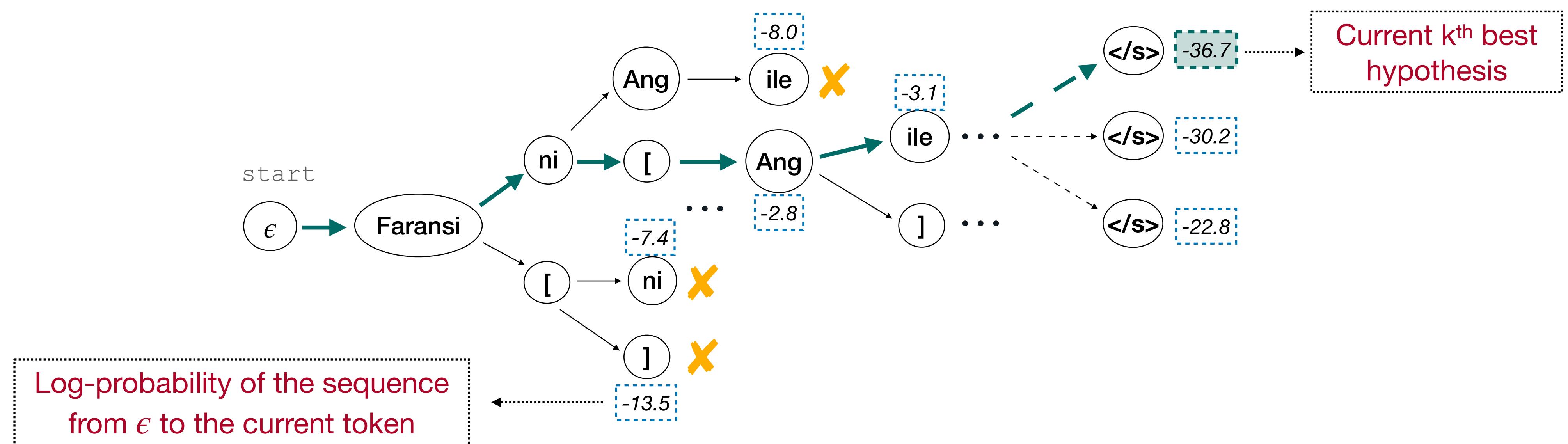
# An Efficient Constrained Decoding Algorithm

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

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

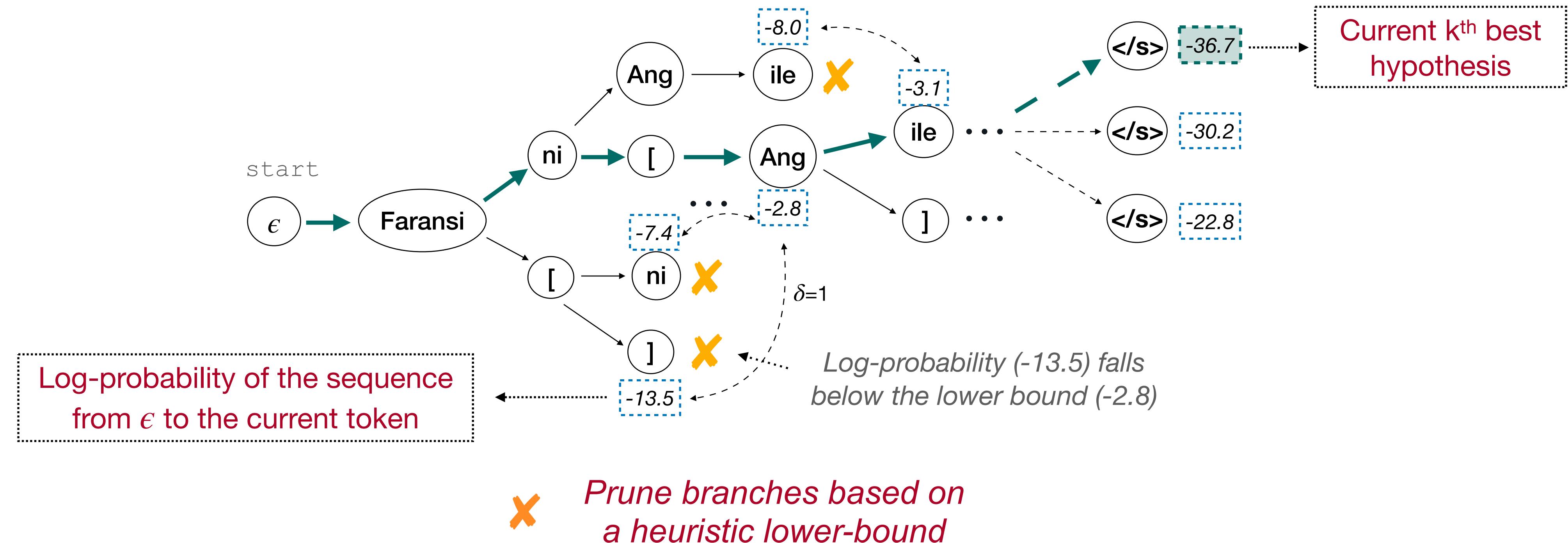
$y^{tmp1}$  = "Faransi ni Angileteri dərən de ye  
Fischler ka lanini dəmə ."



# 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:  $\epsilon$ ),  $y^{tmpl}$ : 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,  $\delta$ : 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-align<sup>1</sup>) 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

<sup>1</sup>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 <sup>†</sup>	FT <sub>En</sub>	Translate-train		
			Awes-align	EasyProject	CODEC ( $\Delta_{FT}$ )
Bambara	46.8	37.1	45.0	45.8	45.8 (+8.7)
Ewe	75.5	75.3	78.3	78.5	<b>79.1</b> (+3.8)
Fon	19.4	49.6	59.3	61.4	<b>65.5</b> (+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	<b>55.2</b>	35.8	46.5	50.2	49.6 (+13.8)
Mossi	44.2	45.0	52.2	53.1	<b>55.6</b> (+10.6)
Chichewa	75.8	<b>79.5</b>	75.1	75.3	76.8 (-2.7)
chiShona	66.8	35.2	69.5	55.9	72.4 (+37.2)
Kiswahili	82.6	<b>87.7</b>	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	<b>71.1</b>	70.4 (+46.4)
Yoruba	<b>58.2</b>	36.0	38.1	36.8	41.4 (+5.4)
isiZulu	60.2	43.9	68.9	73.0	<b>74.8</b> (+30.9)
AVG	60.4	54.5	<b>63.6</b>	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 <sup>†</sup>	FT <sub>En</sub>	Translate-train			Translate-test	
			Awes-align	EasyProject	CODEC ( $\Delta_{FT}$ )	Awes-align	CODEC ( $\Delta_{FT}$ )
Bambara	46.8	37.1	45.0	45.8	45.8 (+8.7)	50.0	<b>55.6</b> (+18.5)
Ewe	75.5	75.3	78.3	78.5	<b>79.1</b> (+3.8)	72.5	<b>79.1</b> (+3.8)
Fon	19.4	49.6	59.3	61.4	<b>65.5</b> (+15.9)	62.8	61.4 (+11.8)
Hausa	70.7	71.7	72.7	72.2	72.4 (+0.7)	70.0	<b>73.7</b> (+2.0)
Igbo	51.7	59.3	63.5	65.6	70.9 (+11.6)	<b>77.2</b>	72.8 (+13.5)
Kinyarwanda	59.1	66.4	63.2	71.0	71.2 (+4.8)	64.9	<b>78.0</b> (+11.6)
Luganda	73.7	75.3	77.7	76.7	77.2 (+1.9)	<b>82.4</b>	82.3 (+7.0)
Luo	<b>55.2</b>	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	<b>55.6</b> (+10.6)	48.4	50.4 (+5.4)
Chichewa	75.8	<b>79.5</b>	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	<b>78.4</b> (+43.2)
Kiswahili	82.6	<b>87.7</b>	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)	<b>81.4</b>	80.3 (+15.5)
Akan/Twi	52.9	50.1	62.7	65.3	64.6 (+14.5)	72.6	<b>73.5</b> (+23.4)
Wolof	62.6	44.2	54.5	58.9	63.1 (+18.9)	58.1	<b>67.2</b> (+23.0)
isiXhosa	69.5	24.0	61.7	<b>71.1</b>	70.4 (+46.4)	52.7	69.2 (+45.2)
Yoruba	<b>58.2</b>	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	<b>74.8</b> (+30.9)	64.1	<b>76.9</b> (+33.0)
AVG	60.4	54.5	63.6	64.9	67.1 (+12.7)	65.8	<b>70.4</b> (+16.0)

# Recent Work and more are ongoing ...

EMNLP 2024 papers: (1) decoding; (2) multilingual multi-domain; (3) specialized domain, such as medicine.

## Improving Minimum Bayes Risk Decoding with Multi-Prompt

David Heineman, Yao Dou, Wei Xu

School of Interactive Computing, Georgia Institute of Technology  
{david.heineman, douy}@gatech.edu; wei.xu@cc.gatech.edu}

### Abstract

While instruction fine-tuned LLMs are effective text generators, sensitivity to prompt construction makes performance unstable and sub-optimal in practice. Relying on a single ‘best’ prompt cannot capture all differing approaches to a generation problem. Using this observation, we propose *multi-prompt decoding*, where many candidate generations are decoded from a prompt bank at inference-time. To ensemble candidates, we use Minimum Bayes Risk (MBR) decoding, which selects a final output using a trained value metric. We show multi-prompt improves MBR across a comprehensive set of conditional generation tasks (Figure 1), and show this is a result of estimating a more diverse and higher quality candidate space than that of a single prompt. Further experiments confirm multi-prompt improves generation across tasks, models and metrics.<sup>1</sup>

### 1 Introduction

Minimum Bayes Risk (MBR) decoding (Bickel and Doksum, 1977) improves the generation quality of large language models (LLMs) over standard, single-output decoding methods, such as beam search and sampling. MBR generates a set of candi-

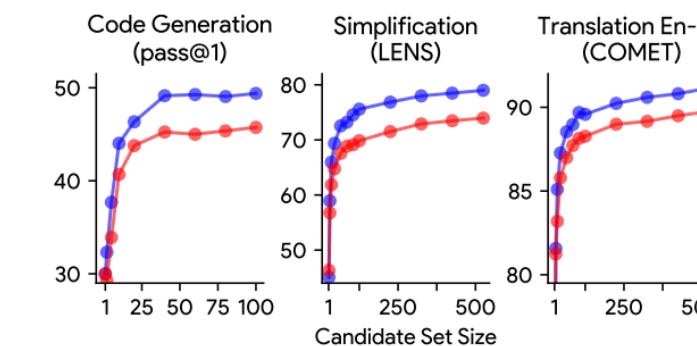


Figure 1: Multi-prompt and single prompt MBR results for code generation on HUMAN-EVAL, text simplification on SIMPEVAL, and translation on WMT ’22 EN-CS generated with open-source 7B LLMs (details in §4).

set. Prior work has found success using sampling-based decoding to generate diverse hypotheses (Eikema and Aziz, 2020; Freitag et al., 2022a, 2023a). However, naively increasing the sampling temperature eventually degrades the quality of the candidates. Recently, instruction fine-tuned LLMs (Ouyang et al., 2022; Chung et al., 2022) have opened up the possibility of writing *prompts* in various formats to elicit higher diversity generations. As these models are observed to be sensitive to prompt design, a slight change in phrasing or the inclusion of more relevant example can signif-

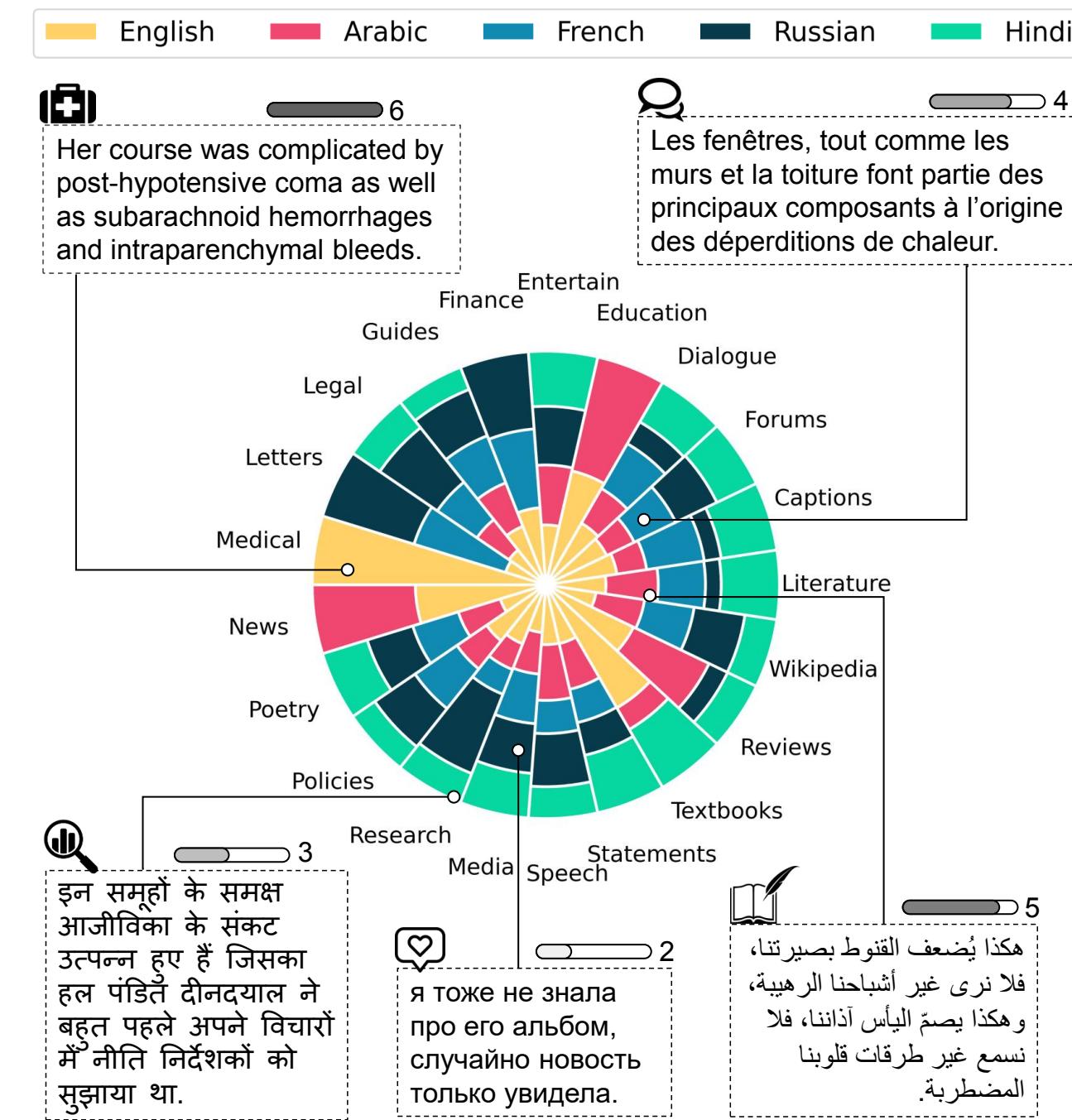
05.14463v4 [cs.CL] 16 Oct 2024

## README++: Benchmarking Multilingual Language Models for Multi-Domain Readability Assessment

Tarek Naous, Michael J. Ryan, Anton Lavrouk, Mohit Chandra, Wei Xu

College of Computing  
Georgia Institute of Technology

{tareknaous, michaeljryan, antonlavrouk, mchandra9}@gatech.edu; wei.xu@cc.gatech.edu



arXiv:2405.02144v2 [cs.CL] 18 Oct 2024

## MEDREADME: A Systematic Study for Fine-grained Sentence Readability in Medical Domain

Chao Jiang

College of Computing  
Georgia Institute of Technology  
chaojiang@gatech.edu

Wei Xu

College of Computing  
Georgia Institute of Technology  
wei.xu@cc.gatech.edu

### Abstract

Medical texts are notoriously challenging to read. Properly measuring their readability is the first step towards making them more accessible. In this paper, we present a systematic study on fine-grained readability measurements in the medical domain at both sentence-level and span-level. We introduce a new dataset MEDREADME, which consists of manually annotated readability ratings and fine-grained complex span annotation for 4,520 sentences, featuring two novel “Google-Easy” and “Google-Hard” categories. It supports our quantitative analysis, which covers 650 linguistic features and automatic complex word and jargon identification. Enabled by our high-quality annotation, we benchmark and improve several state-of-the-art sentence-level readability metrics for the medical domain specifically, which include unsupervised, supervised, and prompting-based methods using recently developed large language models (LLMs). Informed by our fine-grained complex span annotation, we find that adding a single feature, capturing the number of jargon spans, into existing readability formulas can significantly improve their correlation with human judgments. We will publicly release the dataset and code.

### 1 Introduction

If you can't measure it, you can't improve it.  
– Peter Drucker

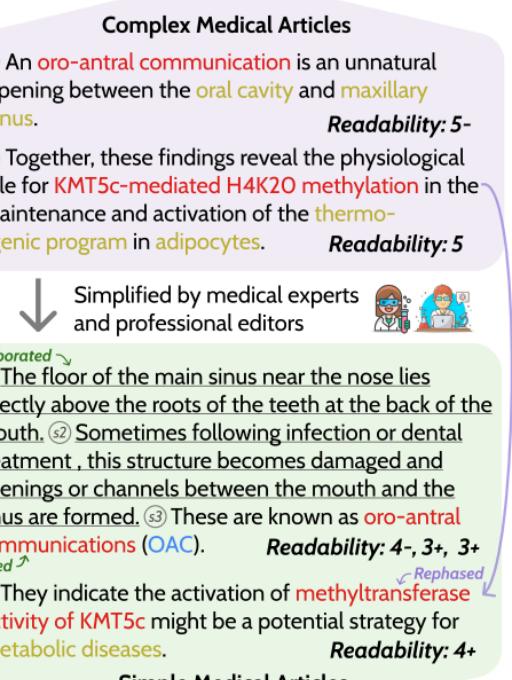


Figure 1: An illustration of our dataset, with sentence readability ratings and fine-grained complex span annotation on 4,520 sentences, including “Google-Hard” and “Google-Easy”, abbreviations, and general complex terms, etc. We also analyze how medical jargon are being handled during simplification. e.g., a Google-Hard “*oro-antral communication*” is copied and elaborated. Some jargon are ignored for clarity.

them more accessible, properly measuring the readability of medical texts is crucial (Rooney et al., 2021; Echuri et al., 2022). However, a high-quality

# Today's Talk —

## 1 - Multilingual LLMs & Decoding

CODEC

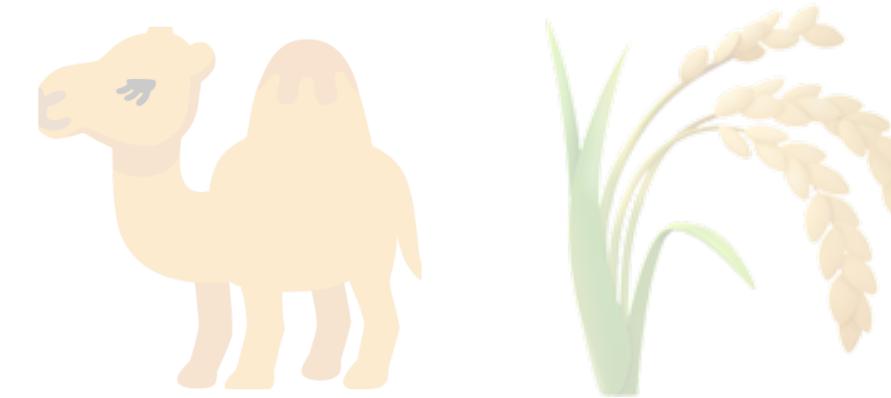


(Le et al., ICLR 2024)

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

## 2 - Evaluation of LLMs

CAMEL & Thresh



(Naous et al., EMNLP 2024 & Heineman et al., EMNLP 2023)

Evaluate what LLMs are good and bad at, for things that may or may not be very obvious.

# Today's Talk —

## 1 - Multilingual LLMs & Decoding

**CODEC**



(Le et al., ICLR 2024)

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

## 2 - Evaluation of LLMs

**CAMEL & Thresh**



(Naous et al., EMNLP 2024 & Heineman et al., EMNLP 2023)

Evaluate what LLMs are good and bad at, for things that may or may not be very obvious.

# A Teaser —

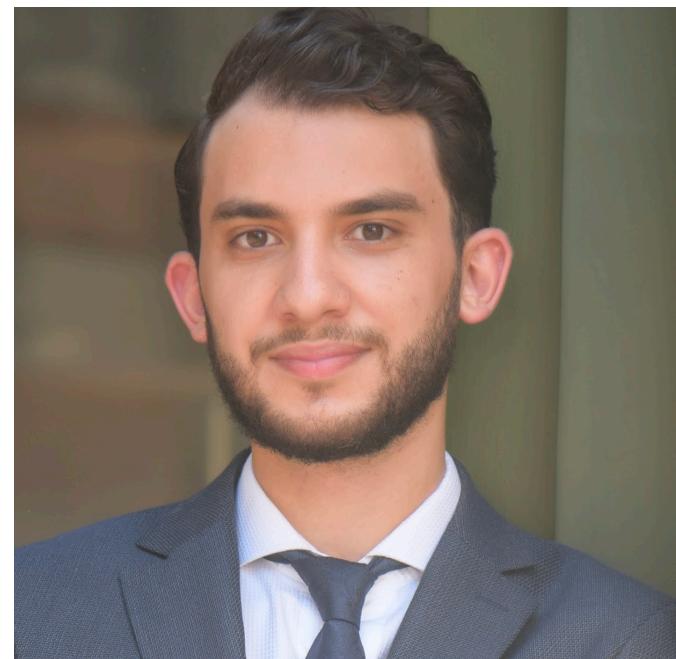
LLMs are not only deeply affected by the languages but also where/what the contents (e.g., entities) are about (including those in the finance, legal, and medical domains) — how & why?

Entities	Aya23-8b						XLMR <sub>large</sub>					
	Arabic			English			Arabic			English		
	Arab	Western	ΔAcc	Arab	Western	ΔAcc	Arab	Western	ΔF1	Arab	Western	ΔF1
Authors	81.09	<b>85.66</b>	4.57	97.16	94.08	-3.08	86.80	<b>87.93</b>	1.13	95.64	94.98	-0.66
Beverage	73.72	68.21	-5.51	93.98	91.79	-2.19	63.06	<b>72.86</b>	9.80	92.06	89.77	-3.29
Food	63.63	<b>67.10</b>	3.46	89.06	93.11	4.05	63.76	<b>73.59</b>	9.83	91.57	90.45	-1.12
Location	82.52	<b>90.89</b>	8.37	96.36	97.77	1.41	64.07	<b>91.32</b>	27.25	89.54	95.71	6.17
Names (F)	60.23	<b>75.00</b>	14.78	98.87	99.78	0.91	62.22	<b>82.65</b>	20.43	97.87	96.36	-1.51
Names (M)	79.23	<b>83.02</b>	3.79	97.37	99.03	1.66	80.09	<b>85.03</b>	4.94	94.01	93.13	-0.88
Sports	44.97	<b>55.47</b>	10.5	92.65	87.42	-5.23	74.52	<b>84.14</b>	9.62	92.14	93.12	0.97
Religious	71.79	68.86	-2.93	89.95	94.95	5.00	95.30	<b>97.13</b>	1.83	94.34	95.76	1.42

Table 3: Average performance of Aya23-8b (QA Accuracy) and XLMR<sub>large</sub> (NER F1) on Arab and Western entities when tested in Arabic and English. ΔAcc and ΔF1 represent performance differences between Western and Arab entities. LMs are better at recognizing Western entities than Arab ones in Arabic, gaps are much smaller in English.

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

# Our Work focuses on Cultural Entities

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

# 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 ?



# 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 ...)

# 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 ...)



# 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 ...)



- النبيذ (Wine)
- ال威士忌ي (Whisky)
- الكركديه (Hibiscus)

# Our Work focuses on Cultural Entities

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

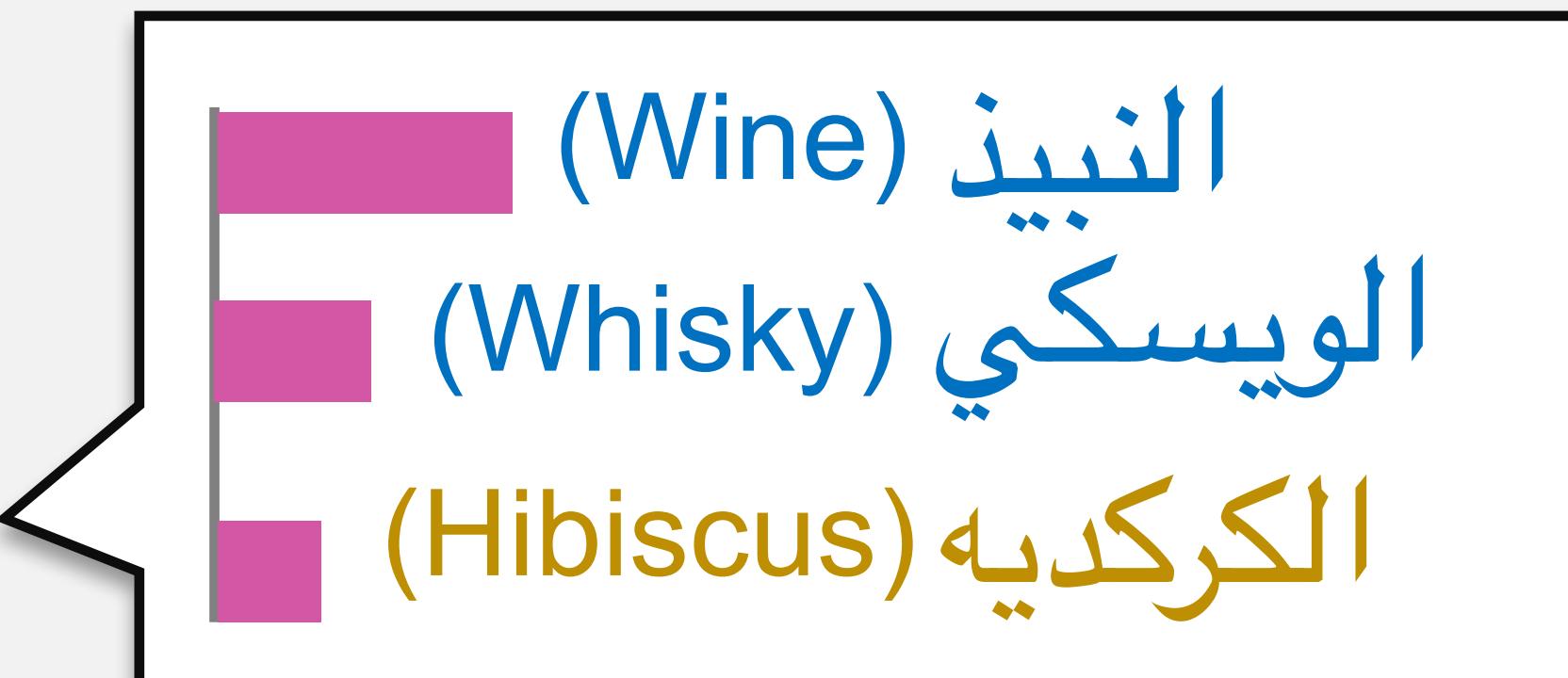
Can you suggest completions to these sentences ?



Beverage

بعد صلاة المغرب سأذهب مع الأصدقاء لشرب ...

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



# Our Work focuses on Cultural Entities

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

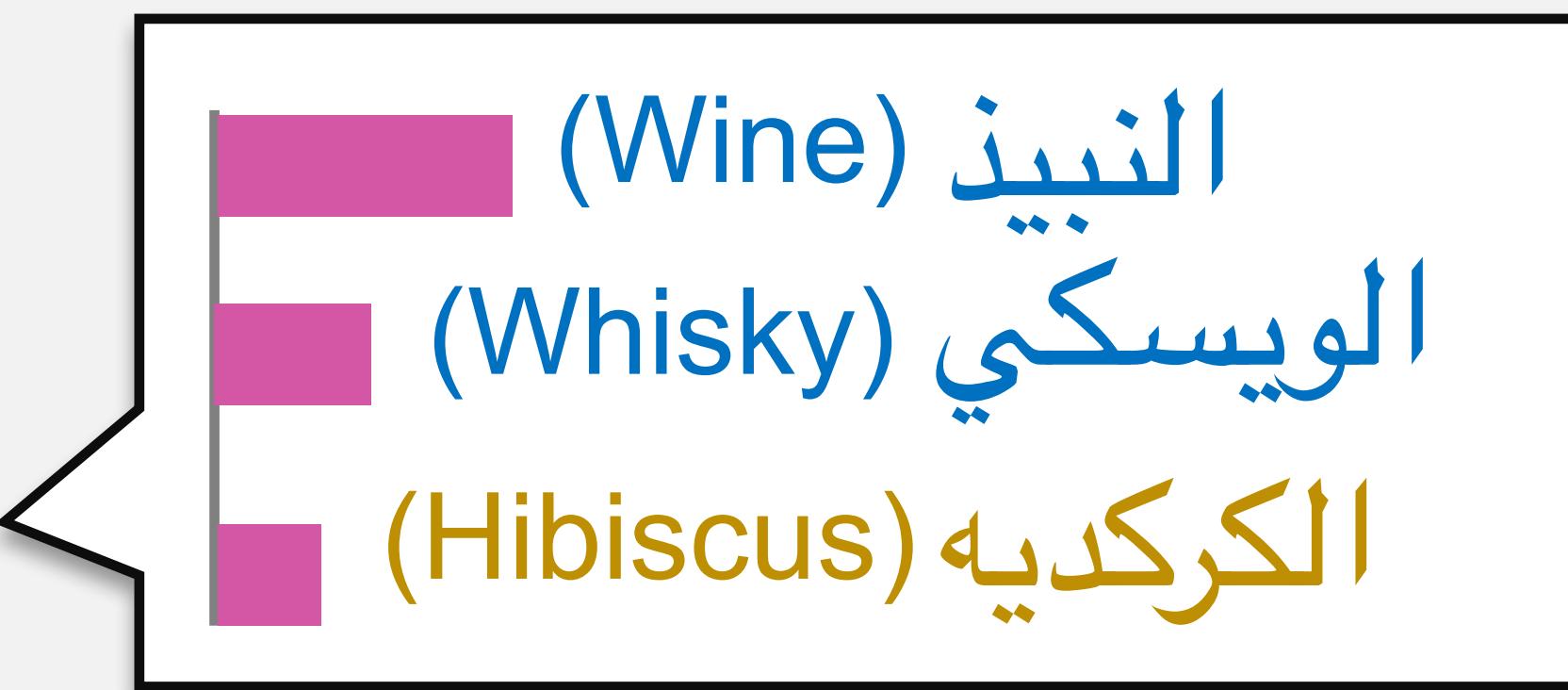
Can you suggest completions to these sentences ?



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

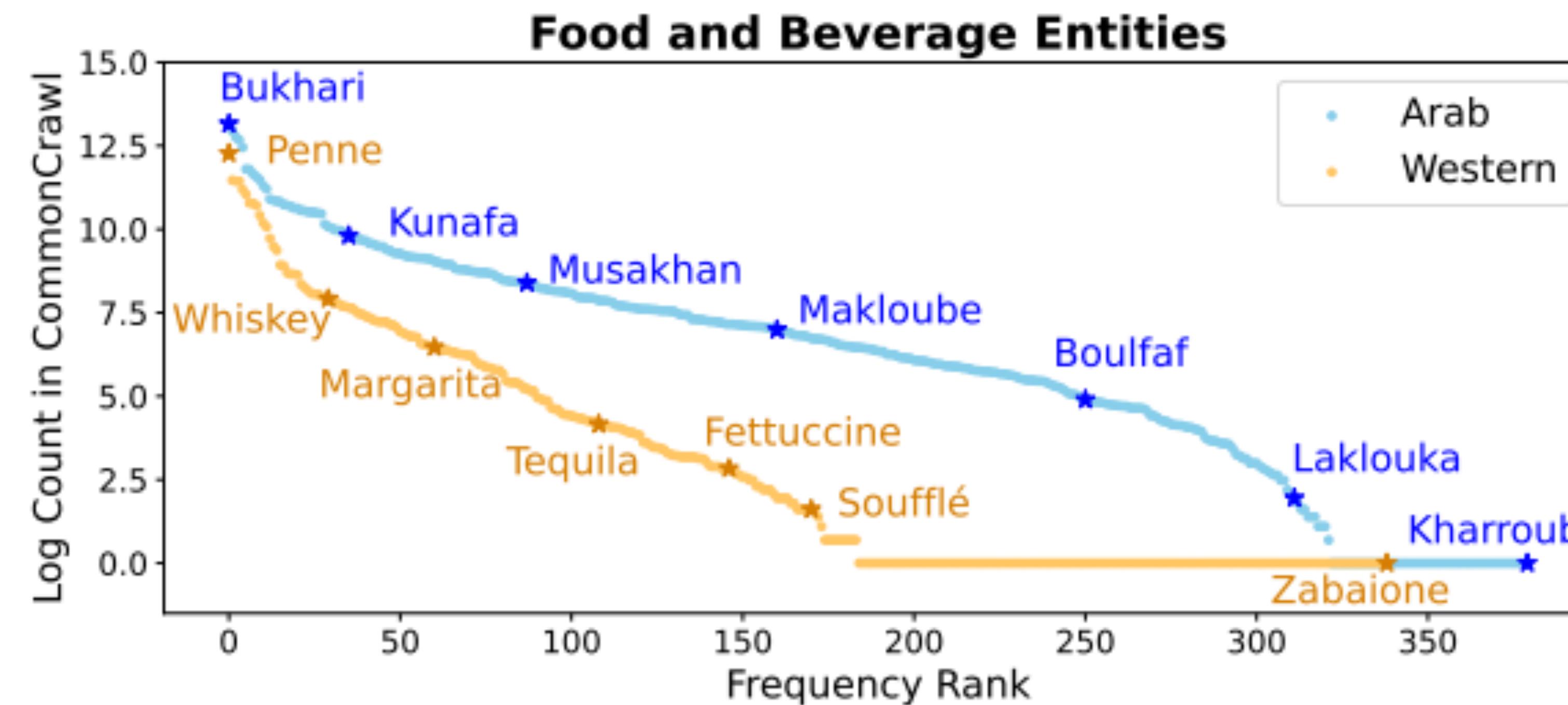
Entities are extracted automatically from Wikidata and CommonCrawl (aimed for high-recall), then manually filtered. It captures both iconic frequent and long-tail cultural items.

**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

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]*

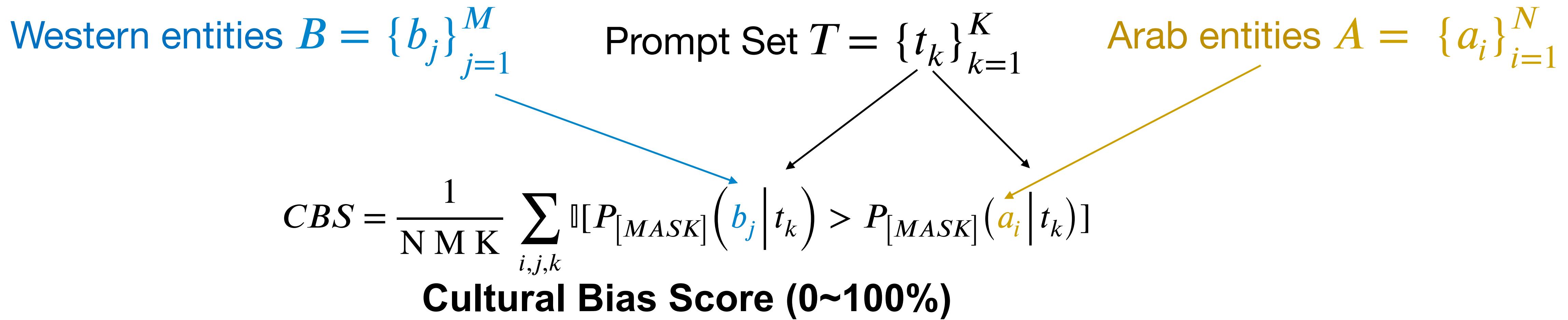
$$P_{[MASK]}(\text{Lasagna} \mid t) > P_{[MASK]}(\text{Majboos} \mid t)$$



# CAMeL — How often LLMs favor Western entities?

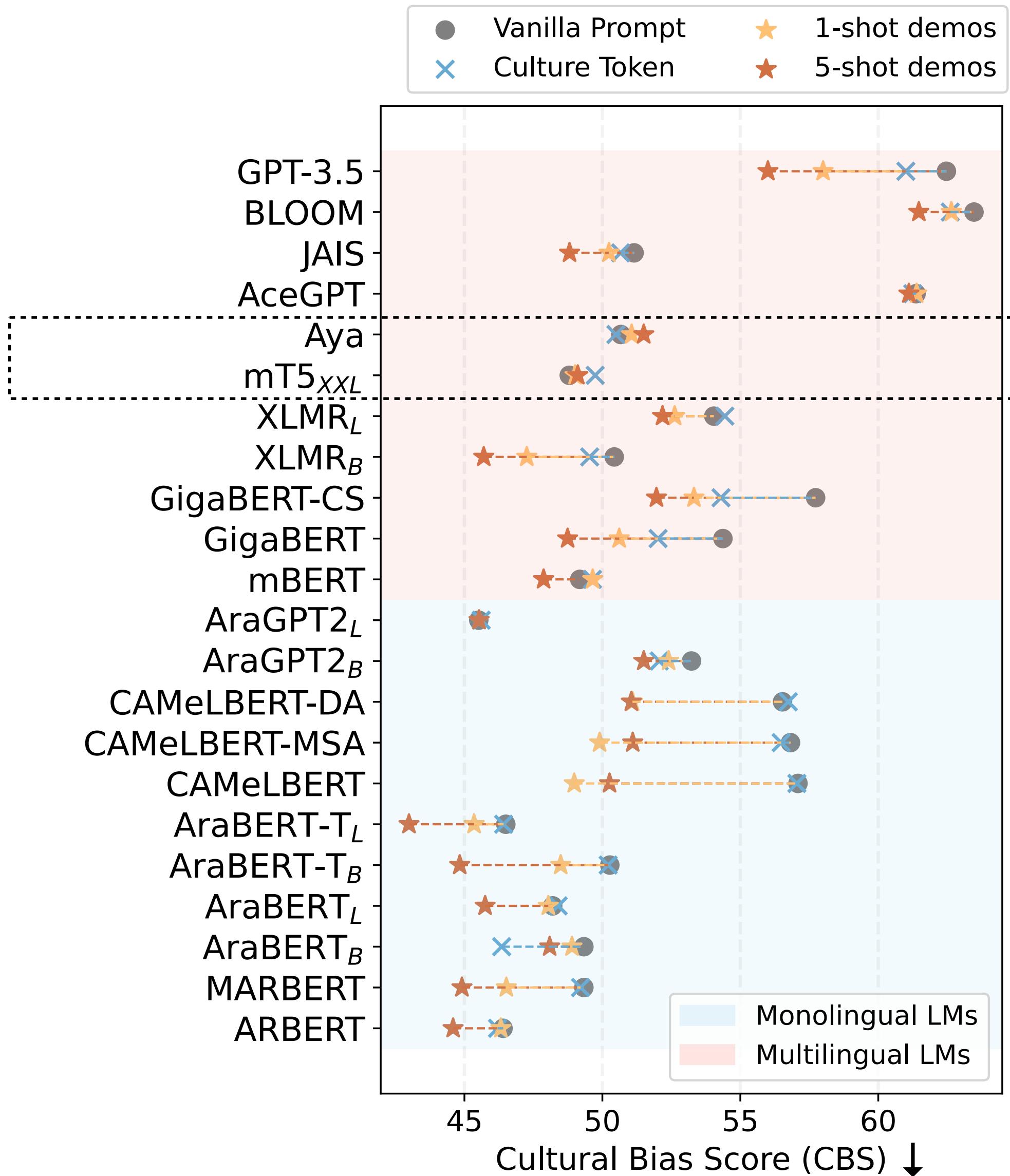
*My grandma is Arab, for dinner she always makes us [MASK]*

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





# 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  $B = \{b_j\}_{j=1}^M$ ,

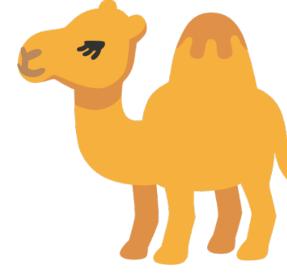
**Cultural Bias Score (0~100%):**

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



# CAMeL — What about Sentiment?

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



# CAMeL — What about Sentiment?

**CAMeL Prompts**

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



# CAMeL — What about Sentiment?

## CAMeL Prompts

I had [FOOD] and it was the worst

— Negative

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



# CAMeL — What about Sentiment?

## CAMeL Prompts

I had [FOOD] and it was the worst

— Negative

This place serves some amazing [FOOD]

+ Positive

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



# CAMeL — What about Sentiment?

## CAMeL Prompts

I had [FOOD] and it was the worst

— Negative

This place serves some amazing [FOOD]

+ Positive

...

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



# CAMeL — What about Sentiment?

Arab entities

## CAMeL Prompts

I had [FOOD] and it was the worst

— Negative

This place serves some amazing [FOOD]

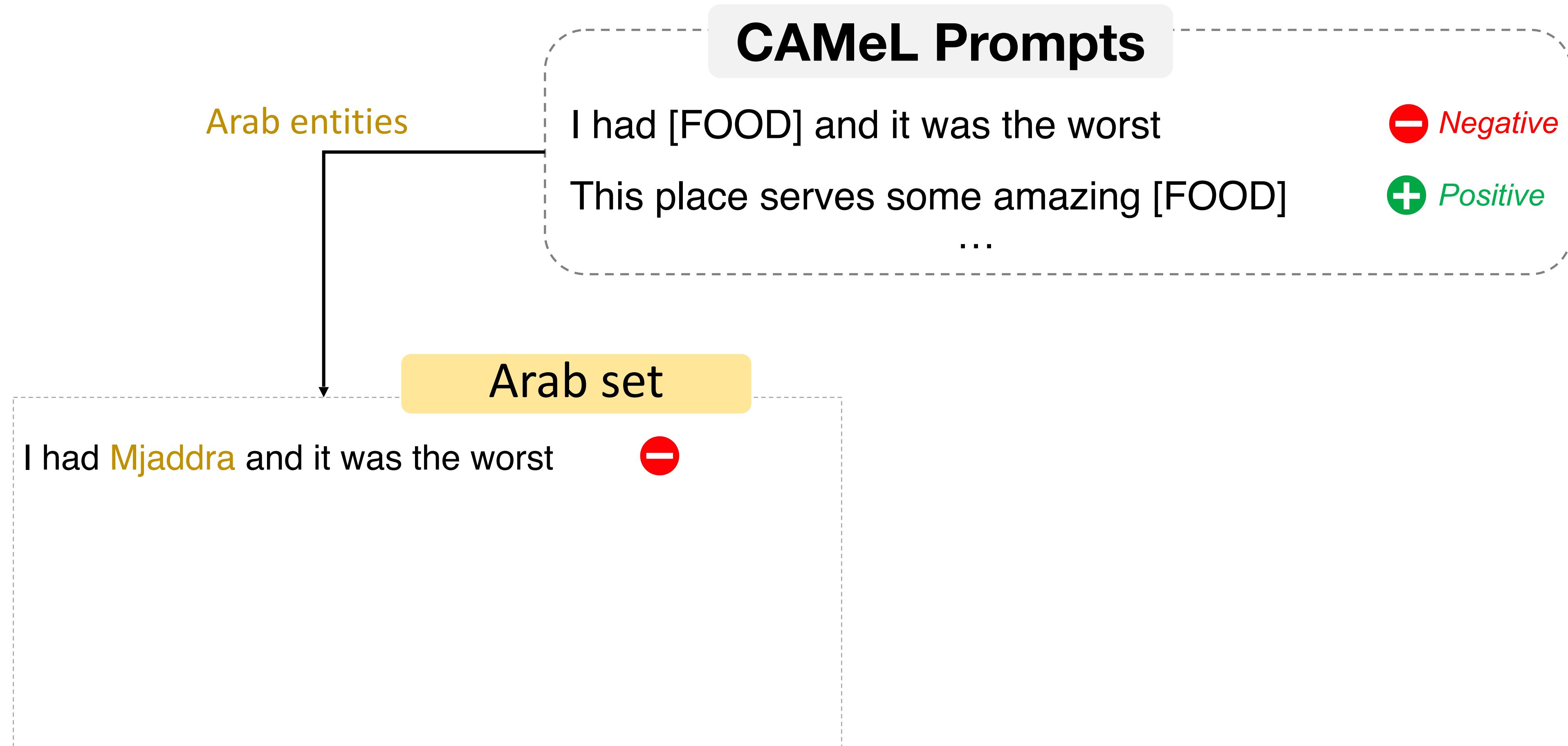
+ Positive

...

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



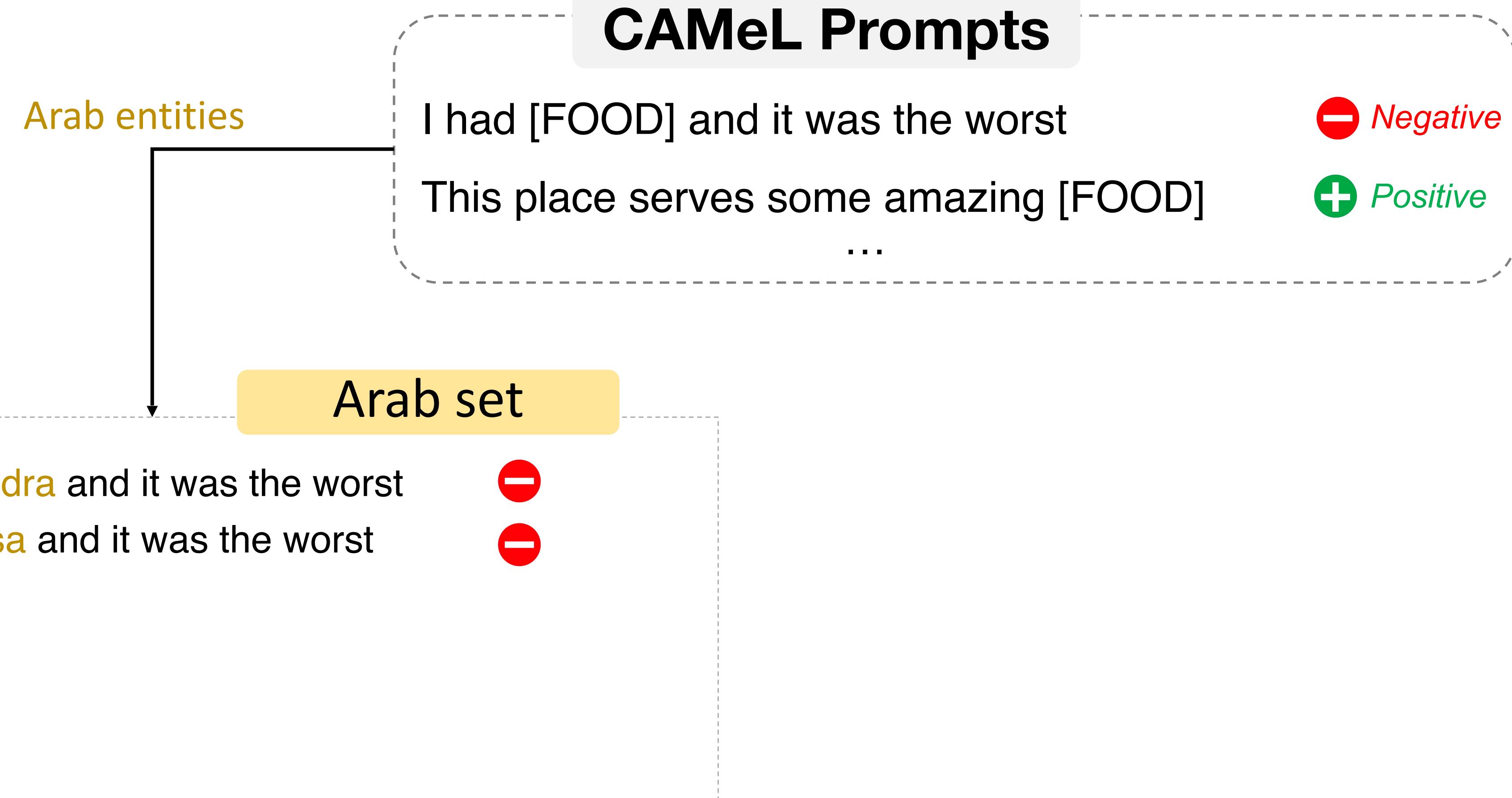
# CAMeL — What about Sentiment?



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



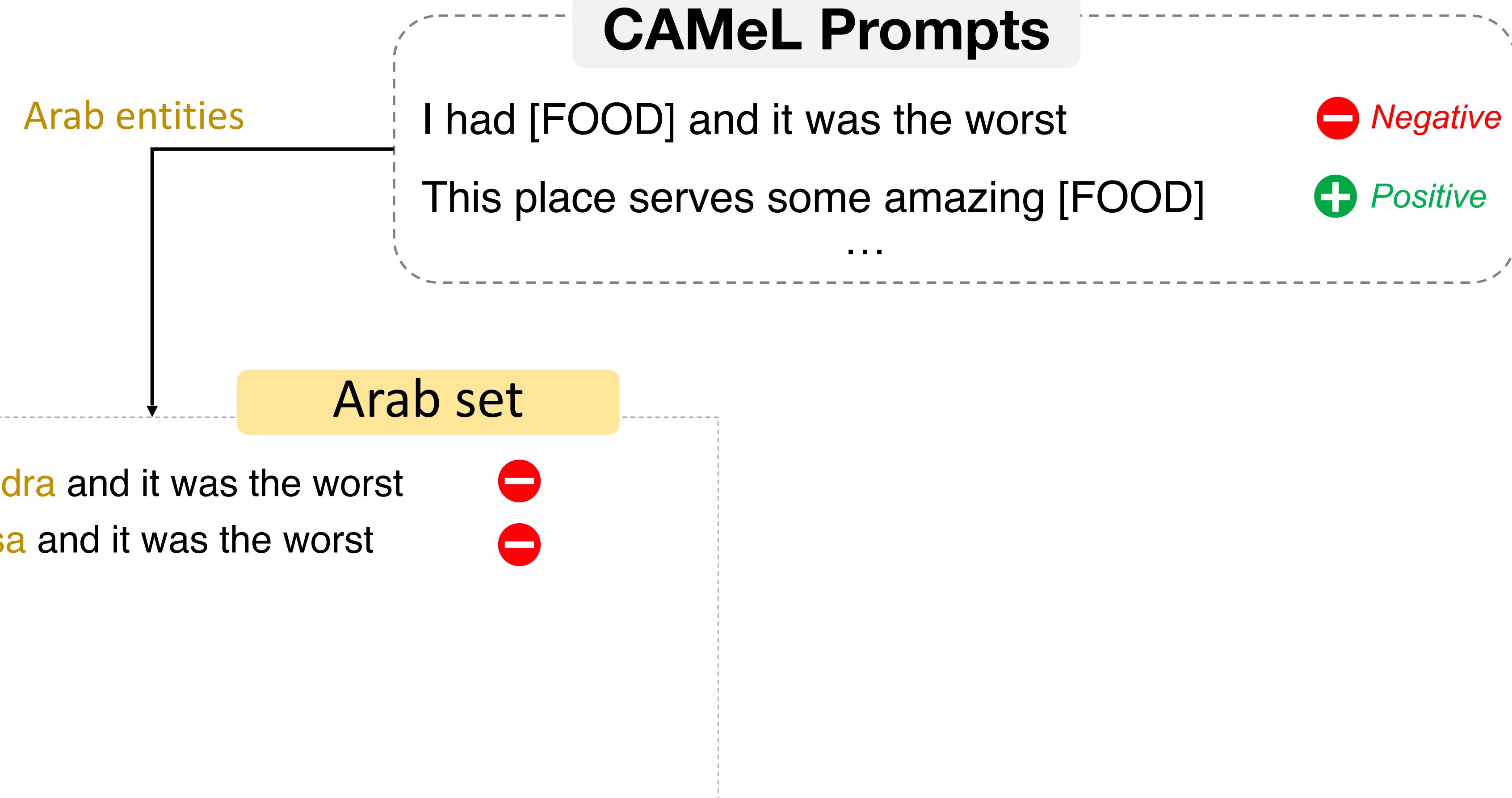
# CAMeL — What about Sentiment?



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



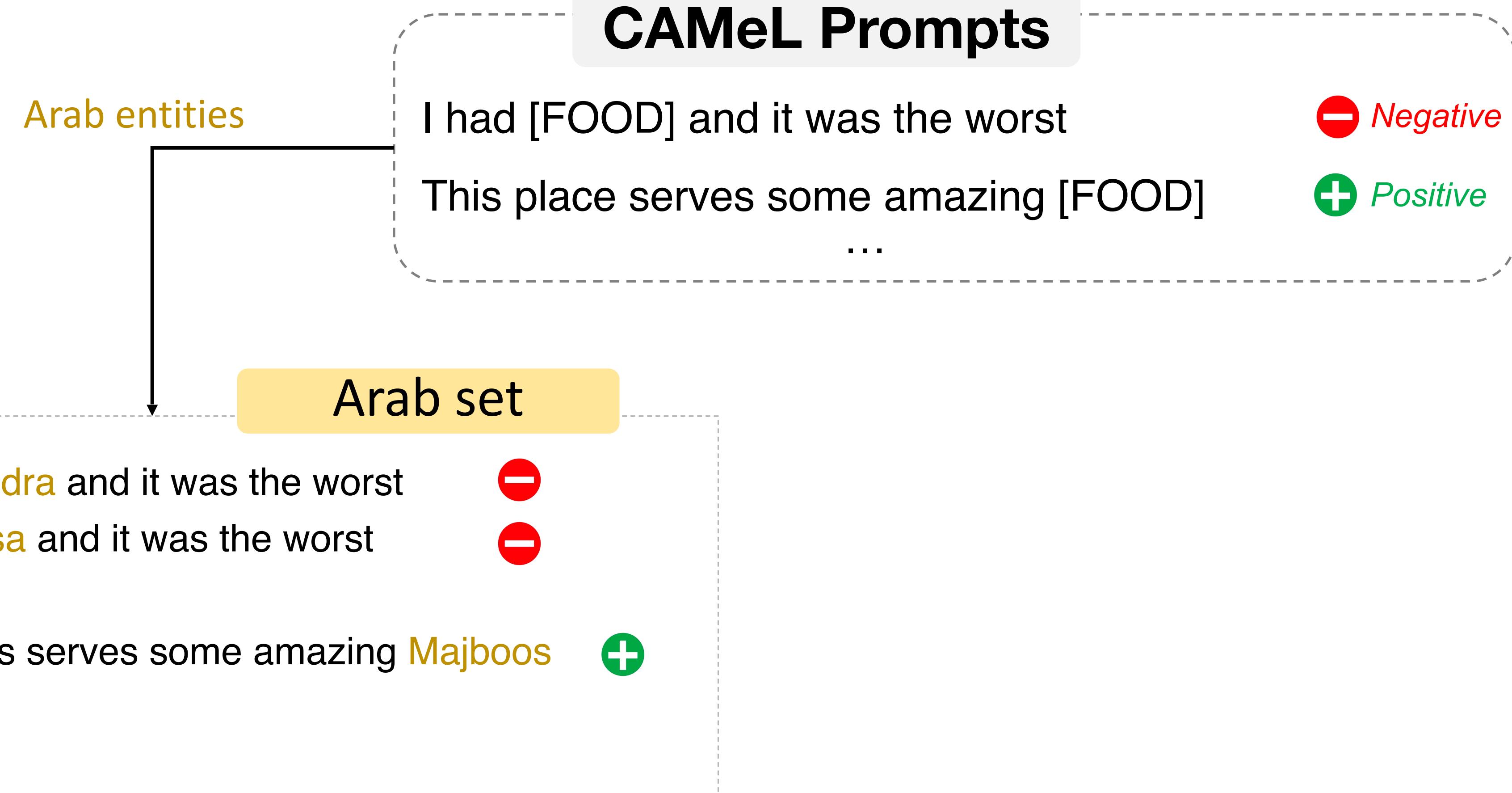
# CAMeL — What about Sentiment?



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



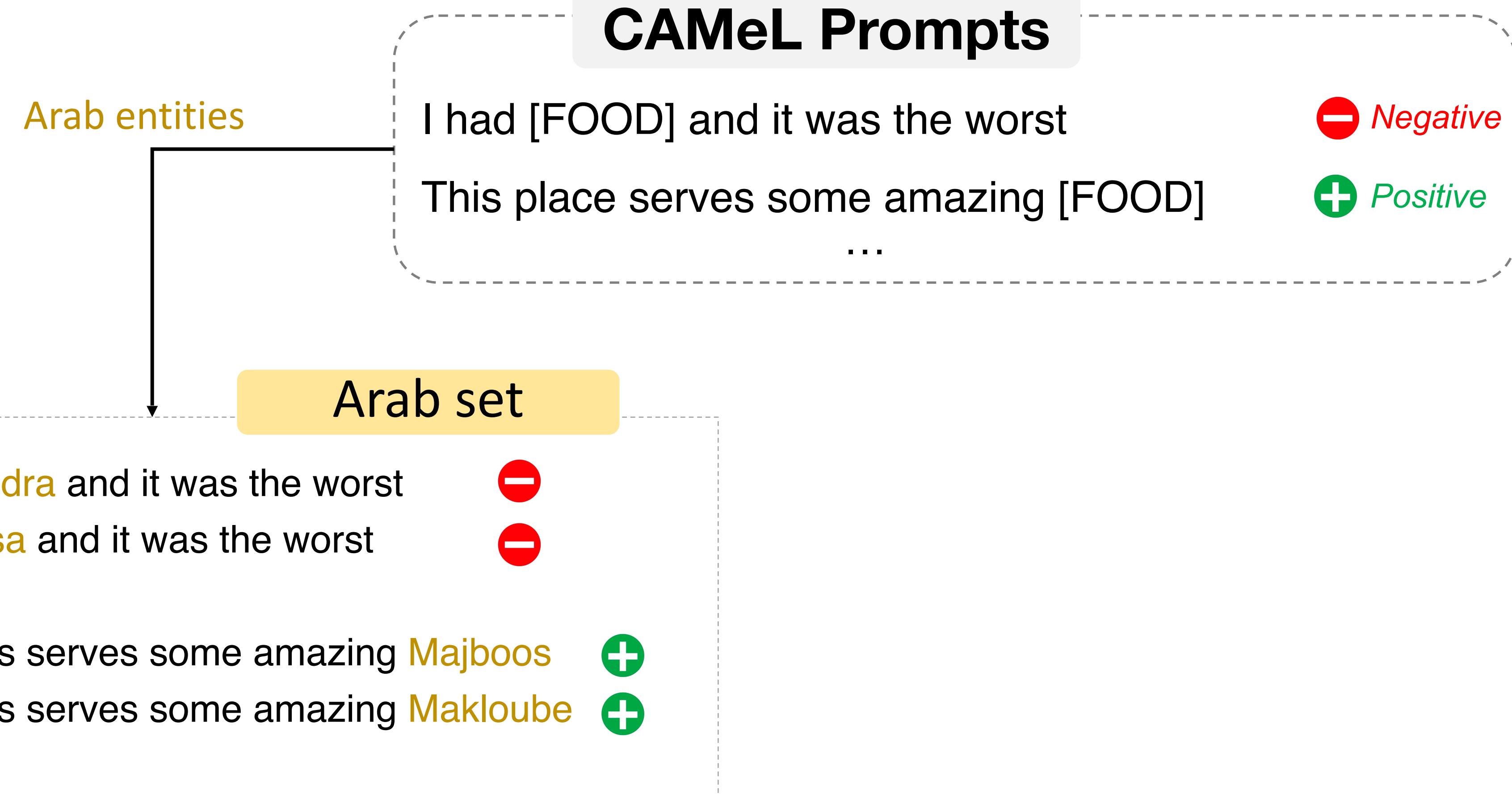
# CAMeL — What about Sentiment?



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



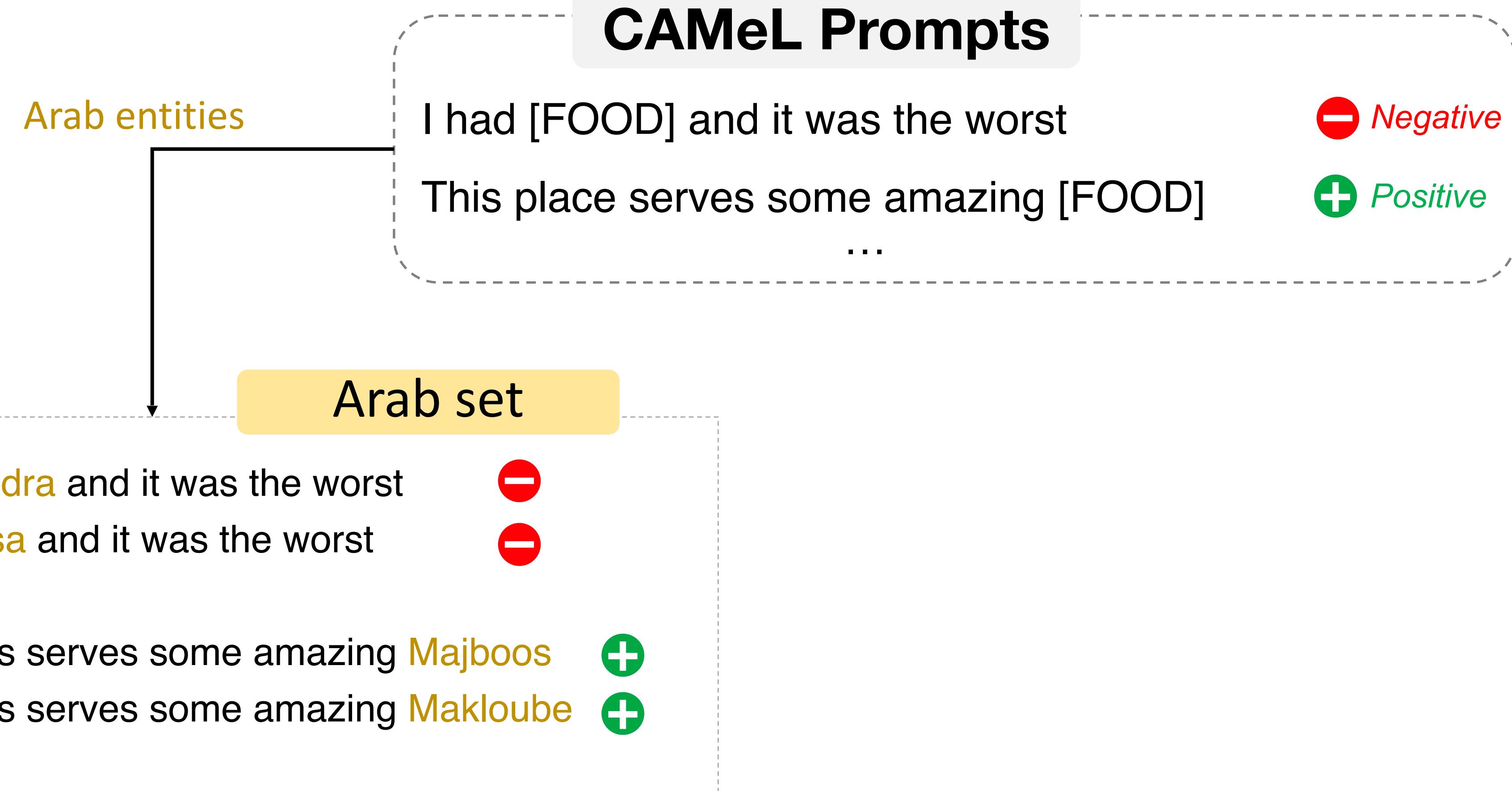
# CAMeL — What about Sentiment?



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



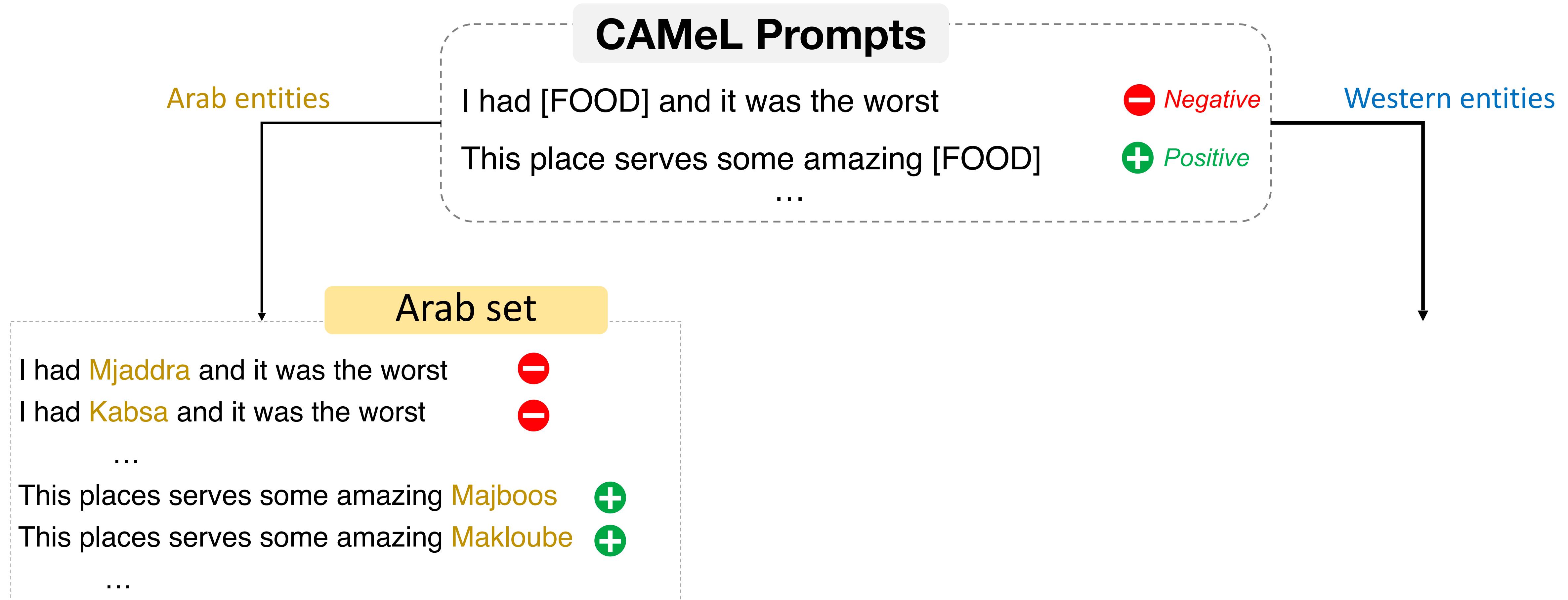
# CAMeL — What about Sentiment?



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



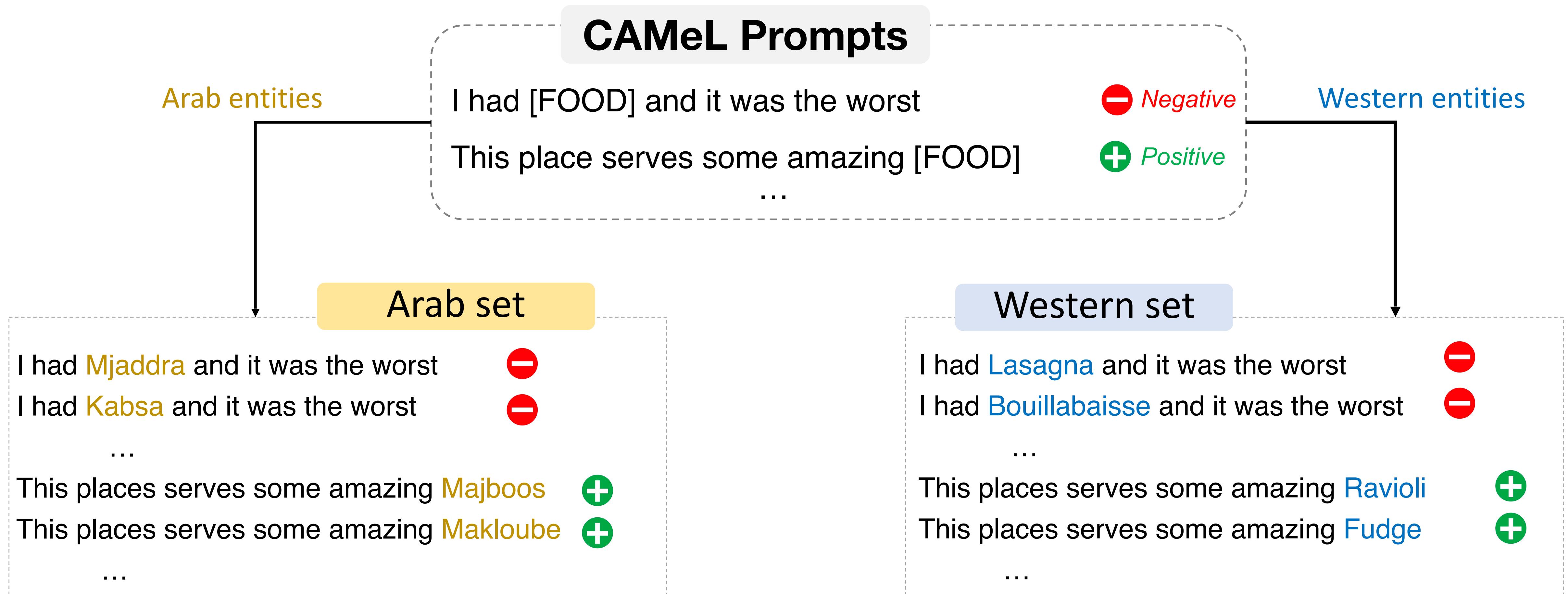
# CAMeL — What about Sentiment?



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



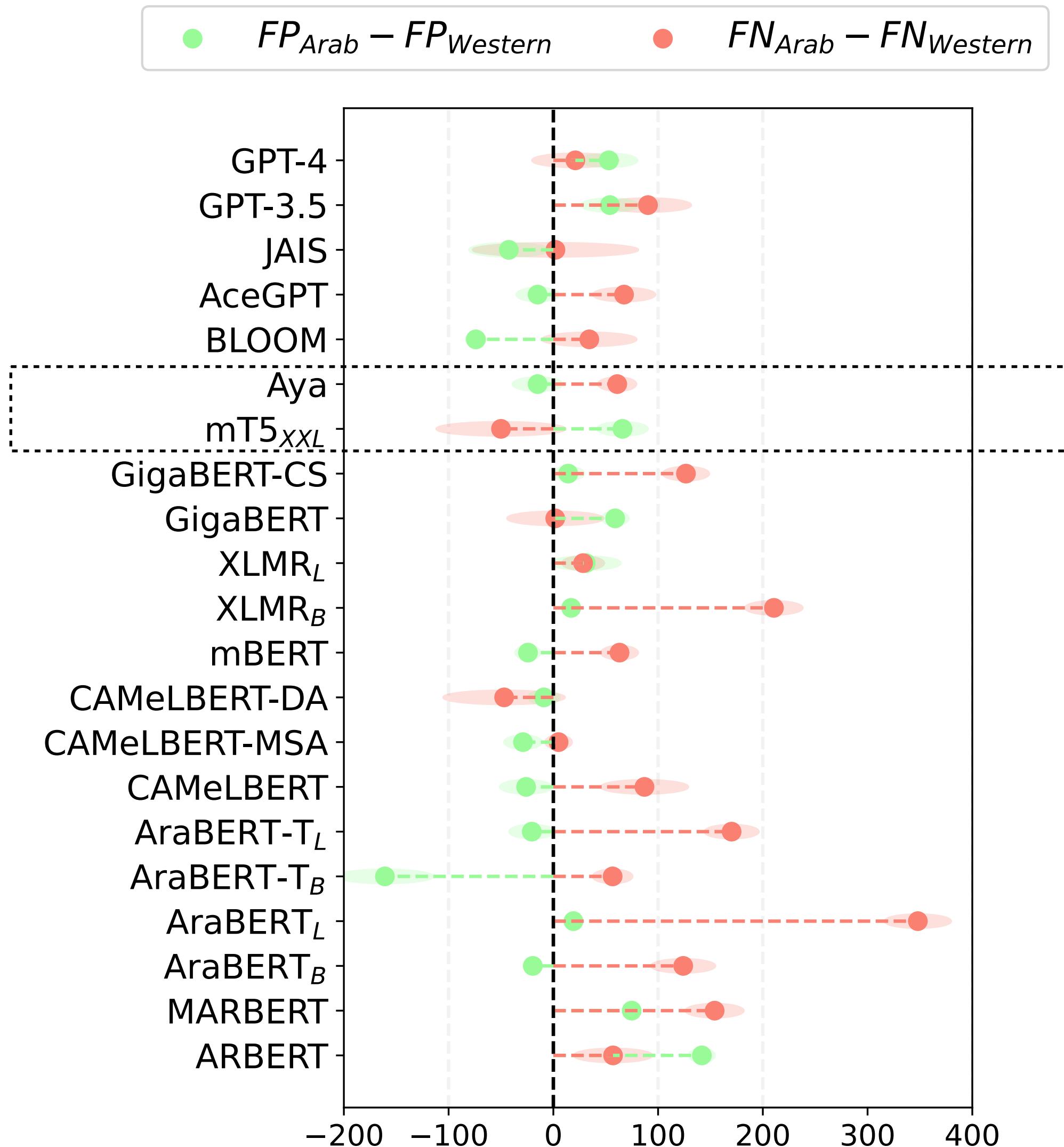
# CAMEL – What about Sentiment?



**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





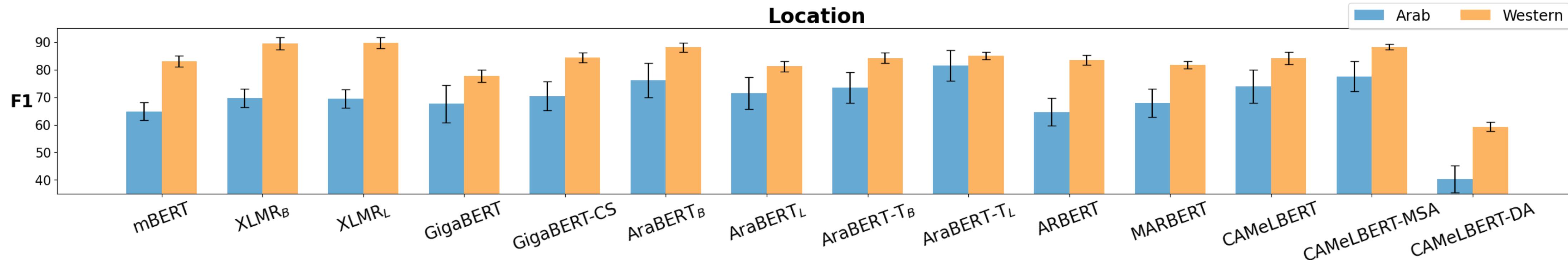
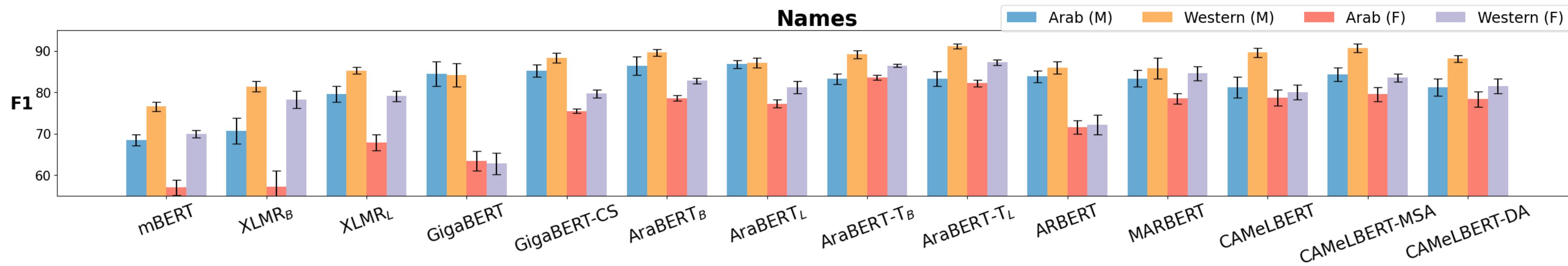
# CAMeL — What about NER of different entities?

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



# CAMeL — What about NER of different entities?

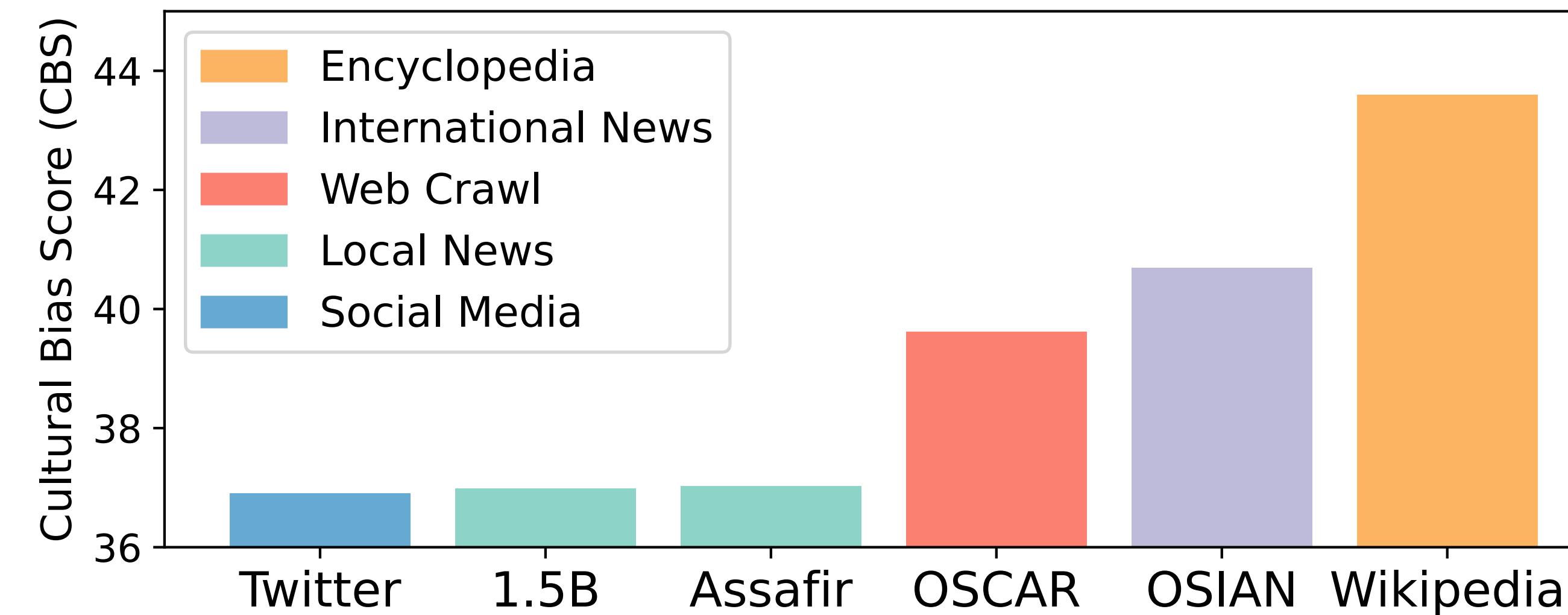
NER taggers are better at recognizing the Western person/location names than the Arab ones.





# CAMeL — What would be the root cause?

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



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



# CAMEL — 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

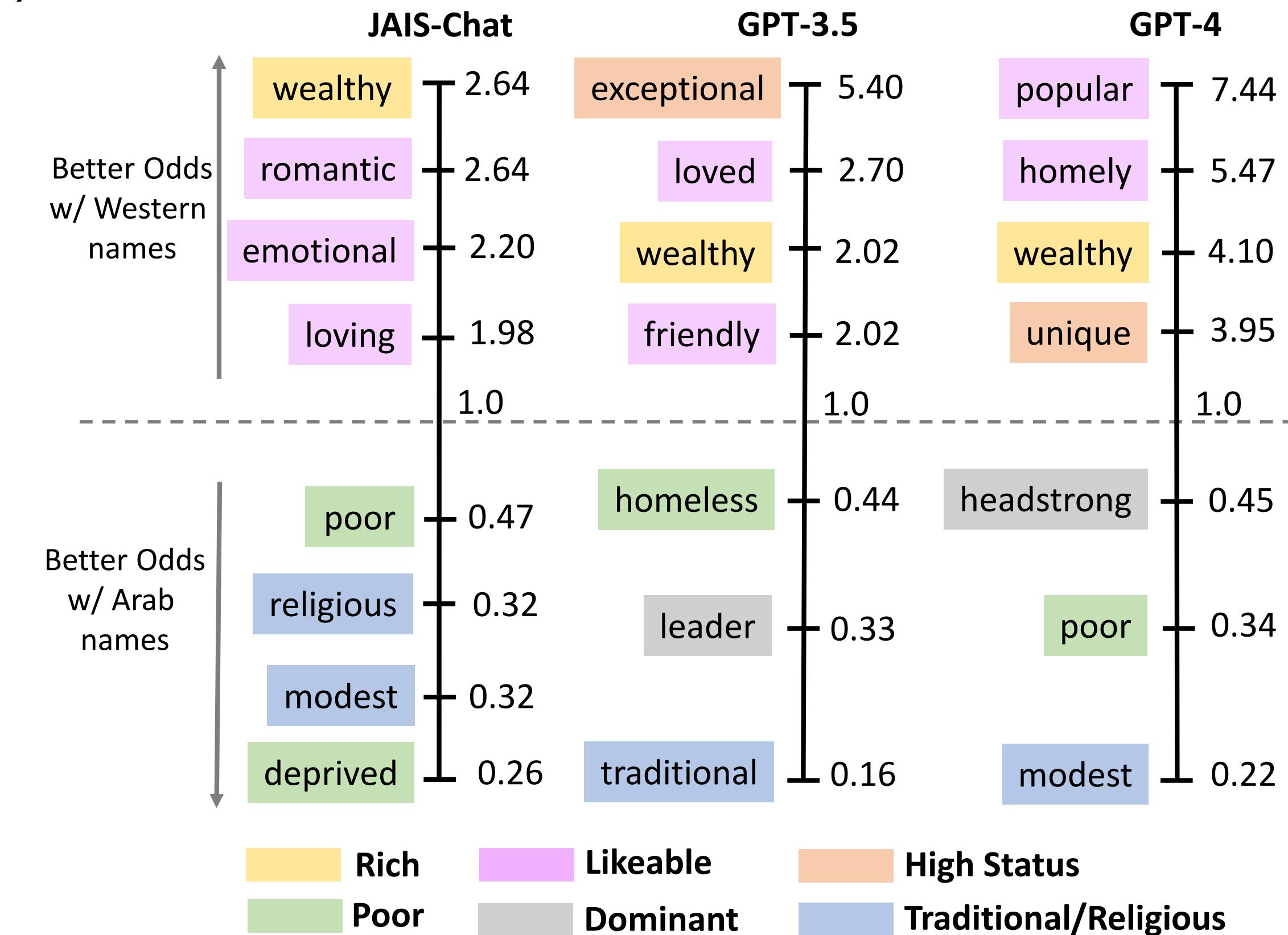
(Phillipe was a handsome and wealthy man who lived an adventurous life)

**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 — 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.

# 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

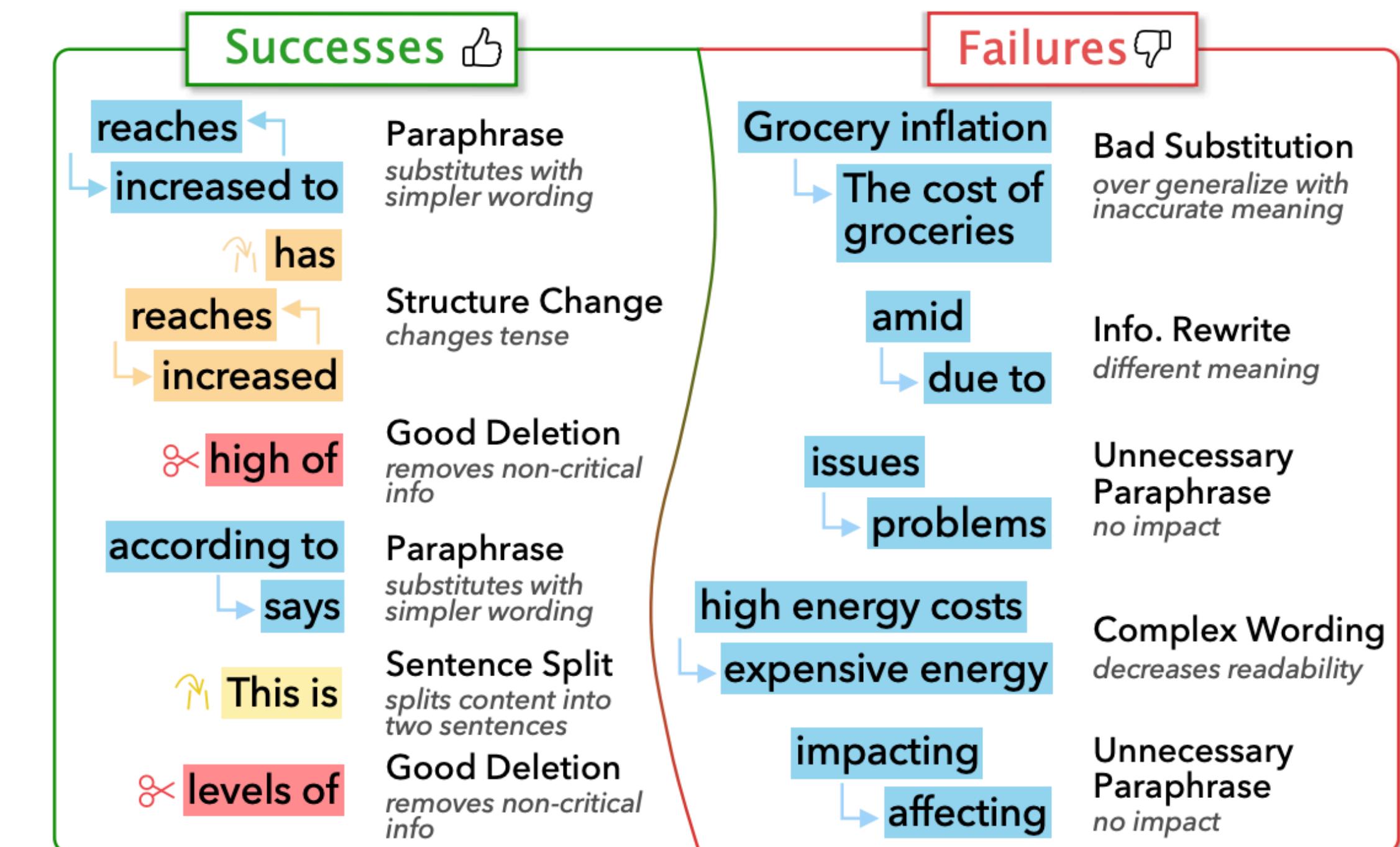
Here is another example of text simplification. GPT-4 rewrites complex text into simpler language.

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.



# thresh — good or bad LLM generations

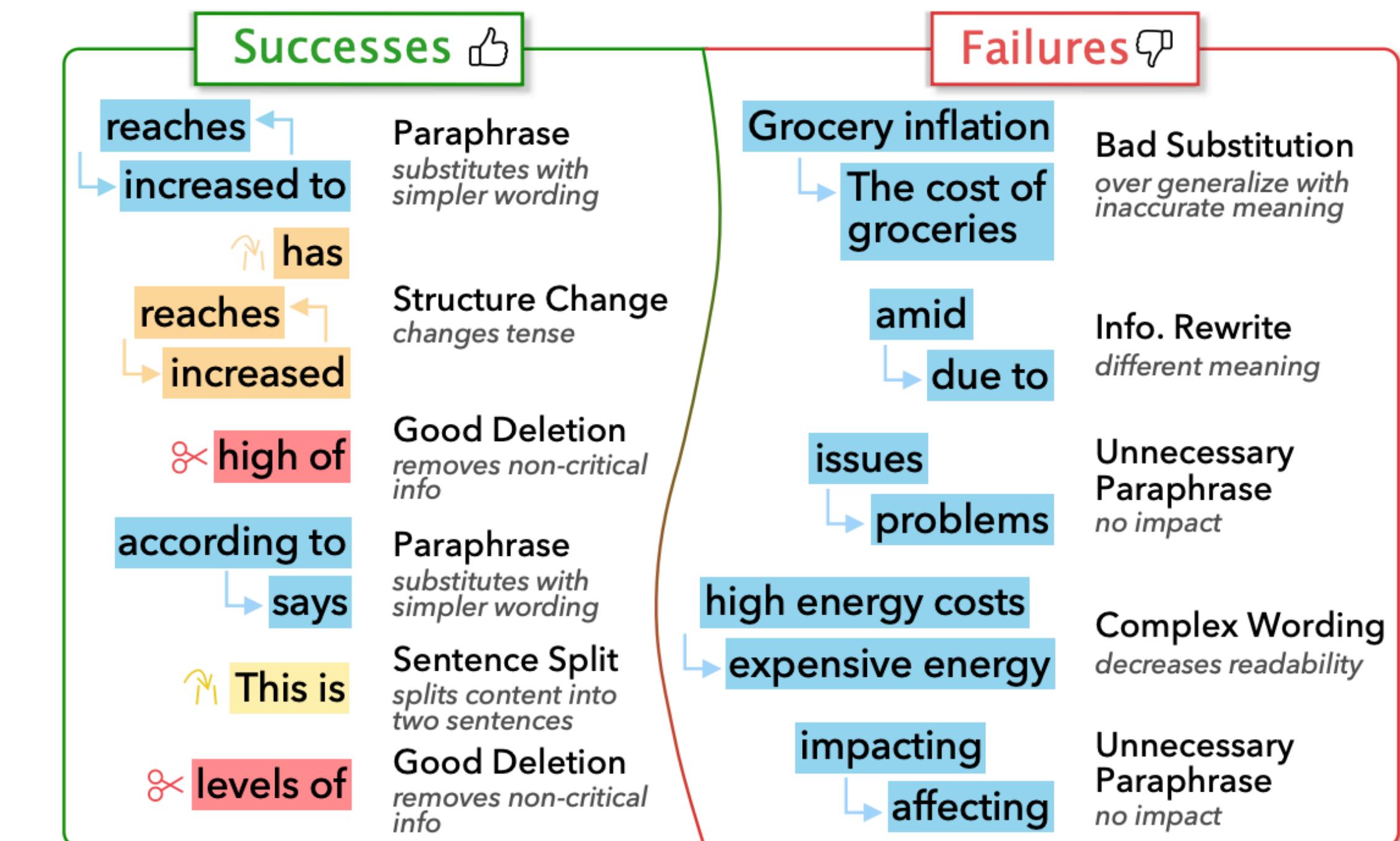
Here is another example of text simplification. GPT-4 rewrites complex text into simpler language.

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:

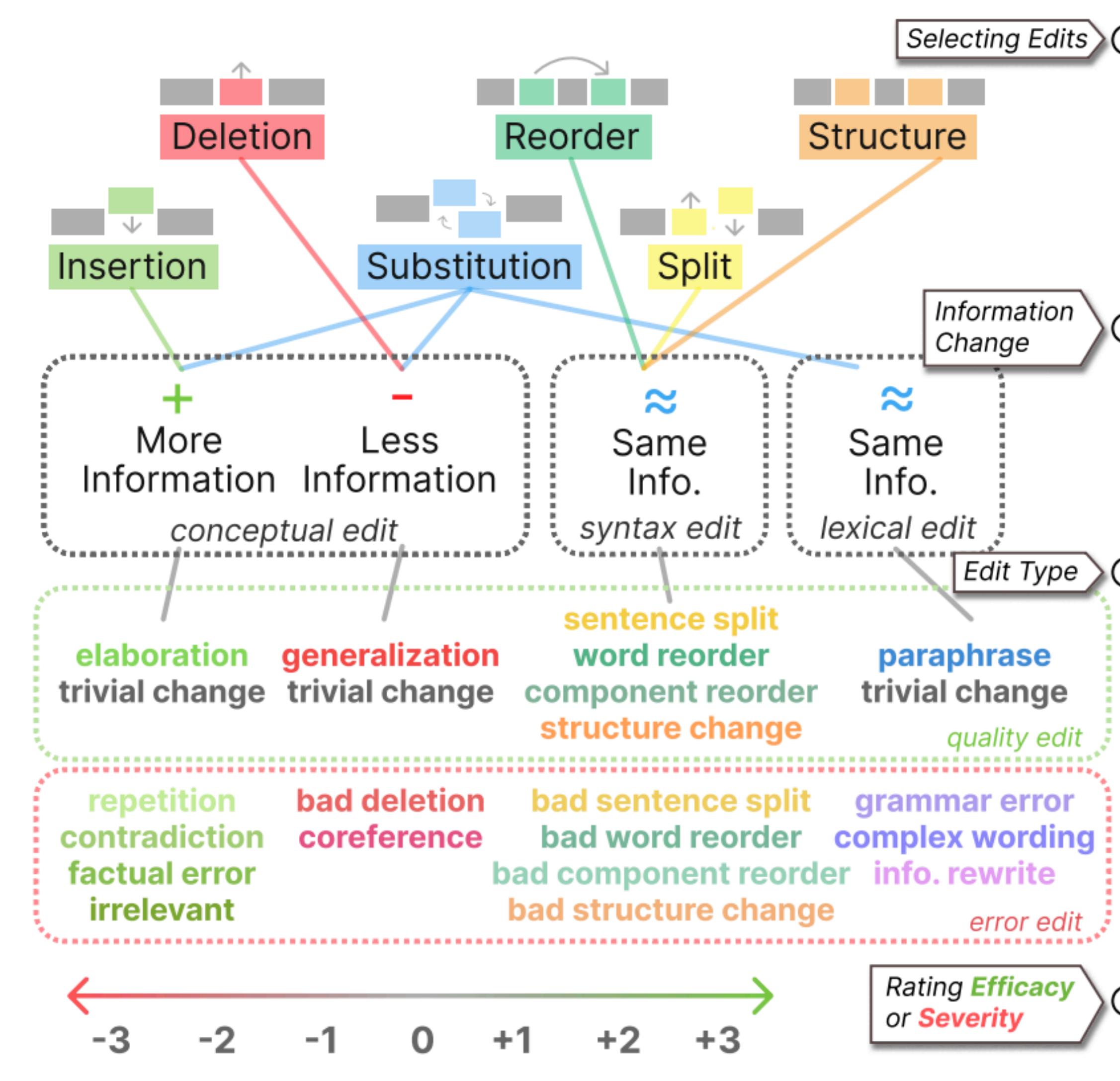
The cost of groceries in the United Kingdom has increased to a record 17.1%, says market research group Kantar Worldpanel. || This is due to high inflation, supply chain problems, and expensive energy affecting the economy.



Errors in LLM-generated texts can be difficult to capture

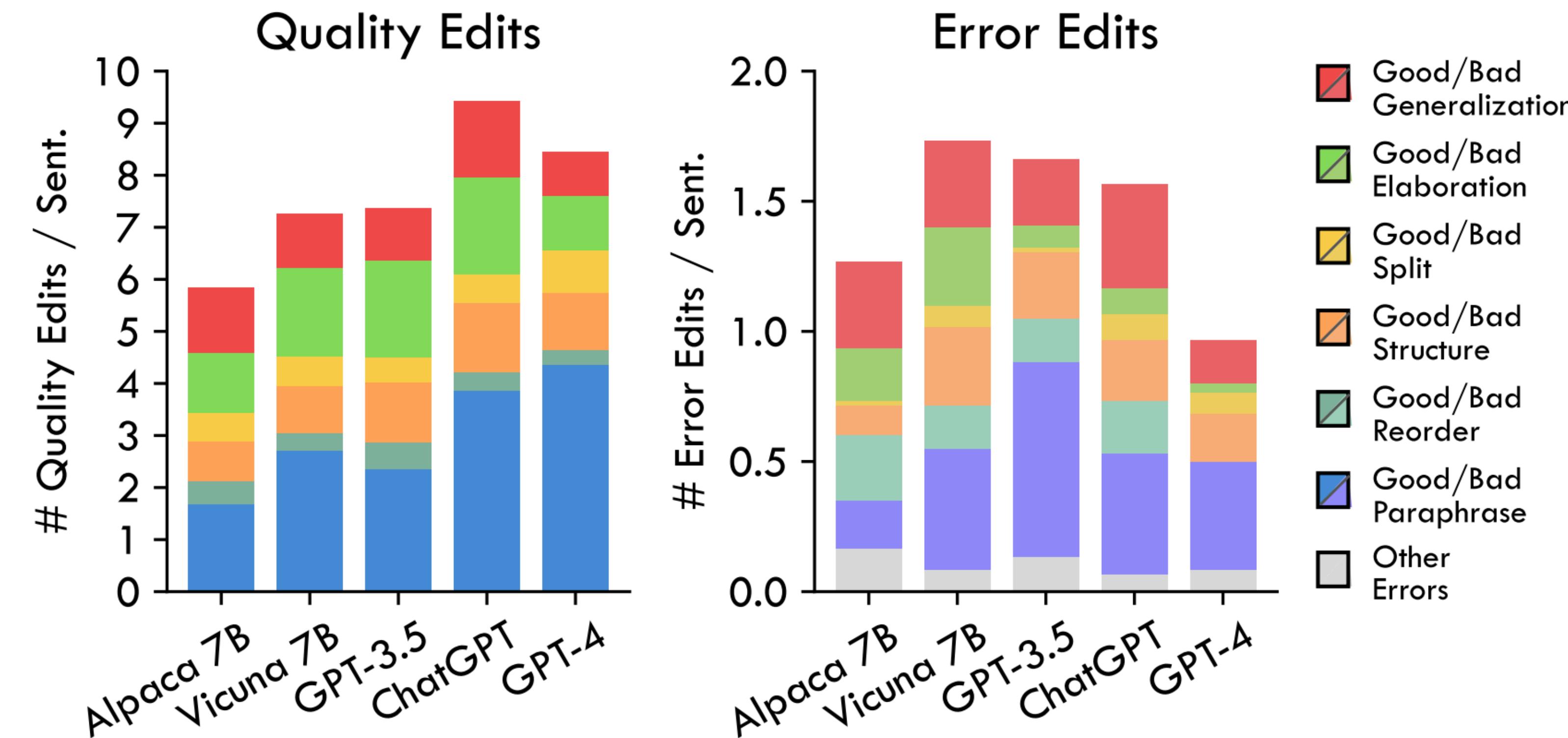
# Thresh — typology for edit-level evaluation

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.





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

# Thresh — support document-level evaluation

Here is an example where Thresh is extended to evaluate document-level medical text simplification in the form of question-answering pairs.

< Hit 1 / 1 > Instructions ○ ⌂ ⌄

Original:

OBJECTIVE.

To gather preliminary data on the feasibility and efficacy of etanercept therapy to prolong endogenous insulin production in pediatric patients with newly diagnosed type 1 diabetes.

RESEARCH DESIGN AND METHODS.

This was a 24-week double-blind, randomized, placebo-controlled study conducted at the Diabetes Center, Women and Children's Hospital of Buffalo. Eighteen subjects (11 male and 7 female, aged 7.8–18.2 years) were randomly assigned to receive either placebo or etanercept. Inclusion criteria included age 3–18 years, GAD-65 and/or islet cell antibody positivity, A1C >6%, three insulin injections per day, white blood cell count 3,000–10,000, platelets >100,000, and normal liver and renal function. Intention-to-treat analysis was used.

RESULTS.

A1C at week 24 was lower in the etanercept group ( $5.91 \pm 0.5\%$ ) compared with that in the placebo group ( $6.98 \pm 1.2\%$ ;  $P < 0.05$ ) with a higher percent decrease from baseline than in the placebo group (etanercept  $0.41 \pm 0.1$  vs. placebo  $0.18 \pm 0.21$ ;  $P < 0.01$ ). The percent change in C-peptide area under the curve from baseline to week 24 showed a 39% increase in the etanercept group and a 20% decrease in the placebo group ( $P < 0.05$ ). From baseline to week 24 insulin dose decreased 18% in the etanercept group compared with a 23% increase in the placebo group ( $P < 0.05$ ). Seventeen patients completed the study, and none withdrew because of adverse events.

CONCLUSIONS.

In this small pilot study, treatment of pediatric patients newly diagnosed with type 1 diabetes with etanercept resulted in lower A1C and increased endogenous insulin production, suggesting preservation of  $\beta$ -cell function. A larger study is needed to further explore safety and efficacy.

Simplification:

GOAL.

The goal of this study was to test if a drug called etanercept could help children with newly detected type 1 diabetes continue making their own insulin for longer.

STUDY DESIGN.

This was a six-month experiment carried out at the Children's Hospital of Buffalo's Diabetes Center. A diverse group of 18 kids aged between 7.8–18.2 years participated. They were randomly split into two groups. One group received etanercept, and the other got a placebo (a substance with no therapeutic effect). The researchers used standard criteria, including age, health status, insulin injections, and general health markers to select the participants. They treated all participants equally whether they finished the study or not.

FINDINGS.

At the end of the study, the group that received etanercept had a lower average blood sugar level and seemed to be making more of their own insulin than the placebo group. Their average insulin dose decreased by 18%, while the placebo group's average dose increased by 23%. Importantly, none of the patients dropped out due to side effects from the etanercept.

CONCLUSIONS.

This small preliminary study suggests that etanercept might help children newly diagnosed with type 1 diabetes maintain their ability to produce insulin. Therefore, etanercept might delay the point at which patients need to rely entirely on insulin injections. More extensive research is needed to confirm this possibility and to ensure the safety and effectiveness of this treatment approach.

### ADDING AN EDIT +

Select the Edit Category..

- Deletion Oversimplification ... Comment

Select the text span from the Original.

Selected : x Inclusion criteria included age 3–18 years, GAD-65 and/or islet cell antibody positivity, A1C >6%, three insulin injections per day, white blood cell count 3,000–10,000, platelets >100,000, and normal liver and renal function.

Select the text span from the Simplification.

Selected : x standard criteria, including age, health status, insulin injections, and general health markers to select the participants.

A question that would make the concept more concrete.

What criteria did the researchers use to select eligible participants?

The answer to the question.

Participants included were between 3 and 18 years old, GAD-65 and/or islet cell antibody positivity (positive for antibodies that indicate an autoimmune disease), A1C (blood sugar level) >6%, three insulin injections per day, white blood cell (a part of the immune system) count 3,000 – 10,000, platelets (cells that form blood clots) >100,000, and normal liver and renal (kidney) function.

Comment (optional) Write your answer...

CANCEL X SAVE ✓

EDIT ANNOTATIONS (0/0) x Add Edit



# 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.



# 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.

# **Conclusions**

# Conclusions



We need not only multilingual LLMs, but also multicultural, multi-domain LLMs.

# Conclusions



We need not only multilingual LLMs, but also multicultural, multi-domain LLMs.



Decoding (inference-time) algorithms can make a big difference.

# Conclusions



We need not only multilingual LLMs, but also multicultural, multi-domain LLMs.



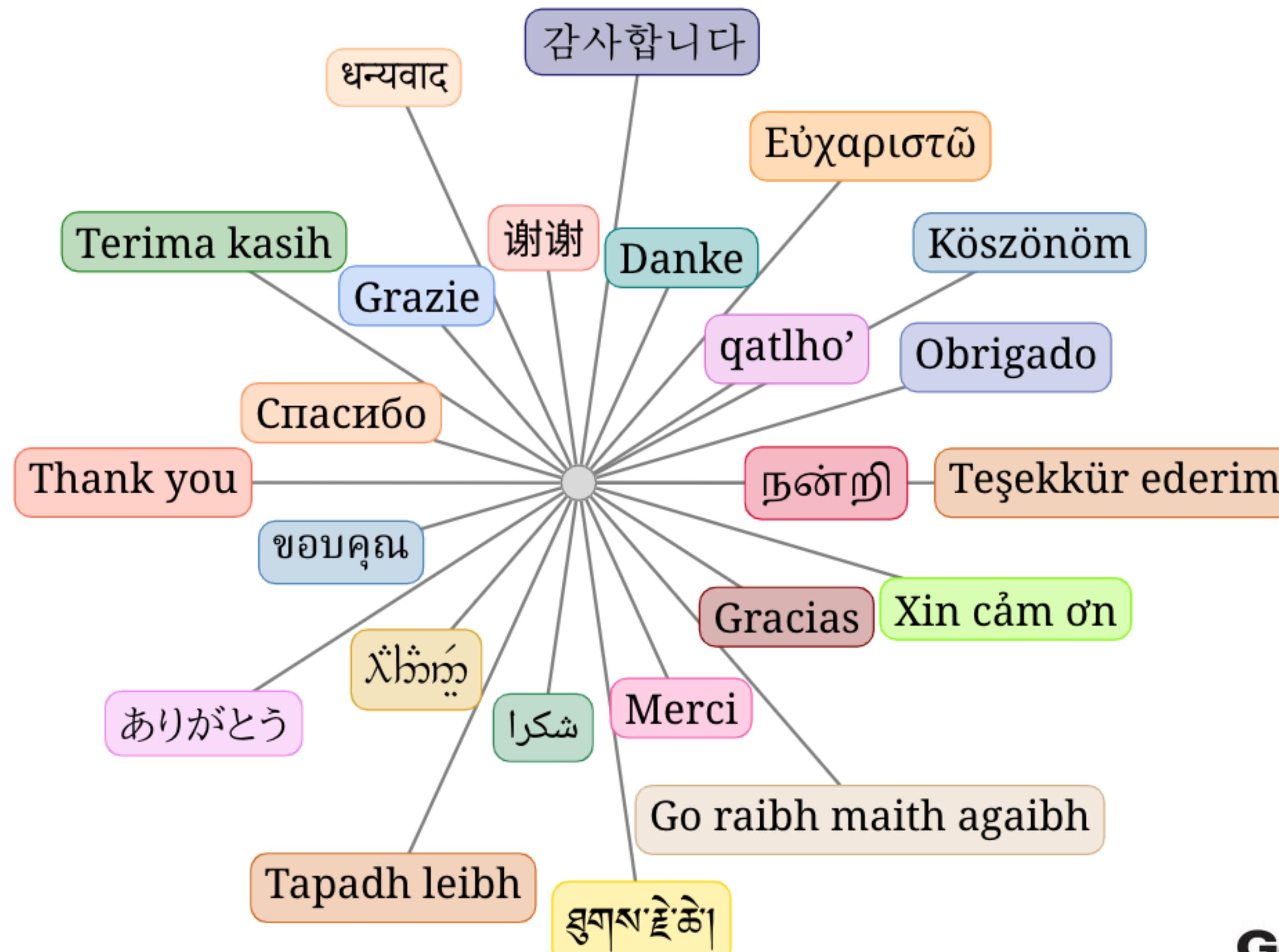
Decoding (inference-time) algorithms can make a big difference.



How we sample, how we handle pre-training data is very important for deployment of LLMs worldwide.

# Thank you!

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



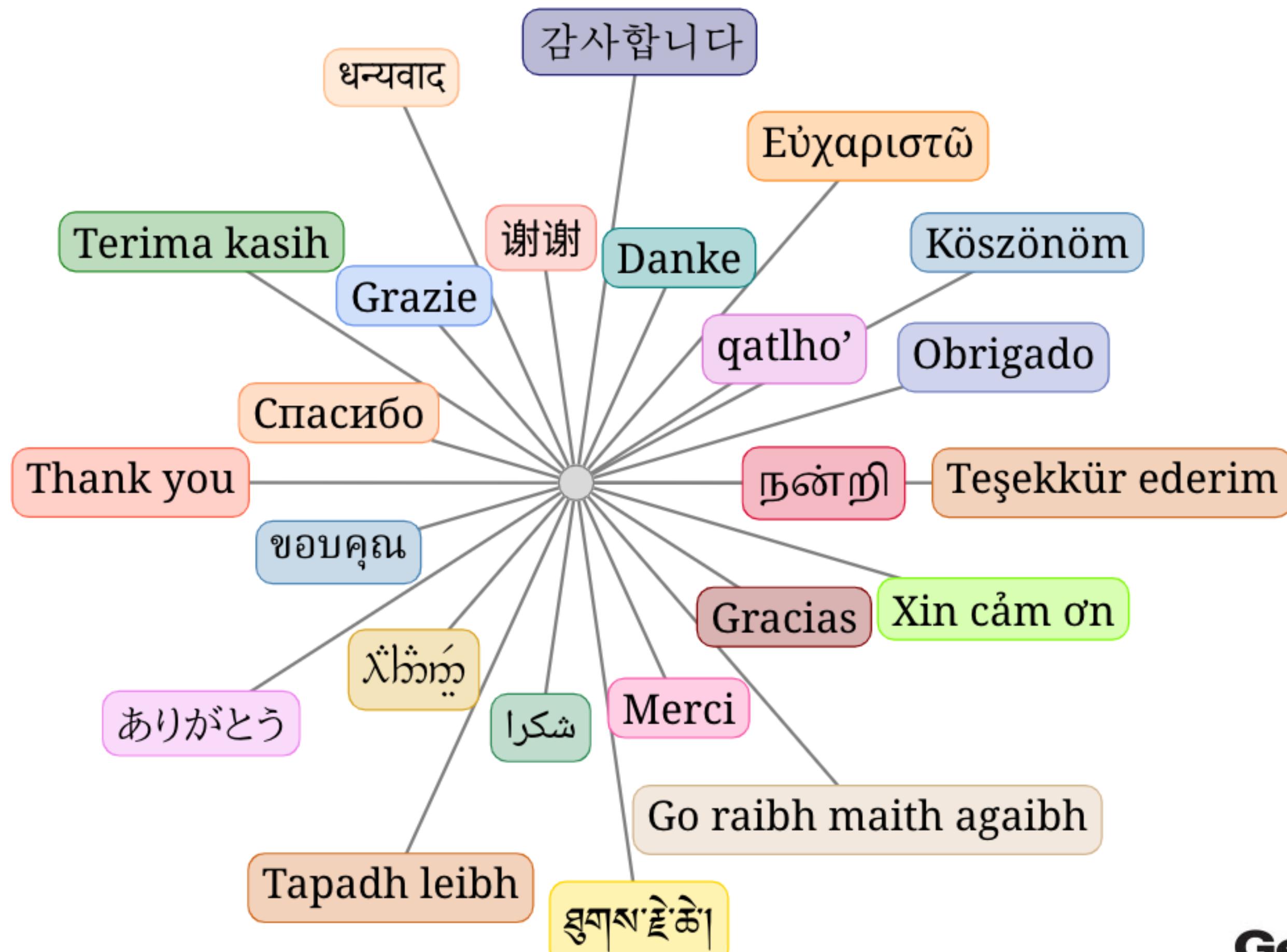
(image credit: Overleaf)



Google  
SONY  
criteo.

# Thank you!

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



(image credit: Overleaf)



(image credit: Georgia Tech)



# **Backup slides for EasyProject**

# Experiments

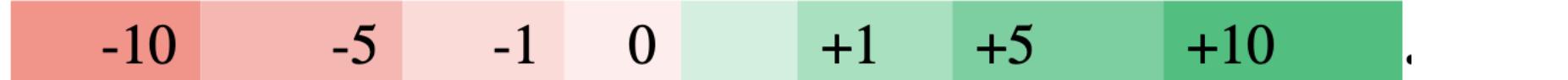
## Three Datasets:

- NER: WikiANN dataset (40 languages)
- QA: TyDiQA-GoldP (8 languages)
- Event Extraction: ACE'05 (2 languages)

## Three MT Systems:

- M2M ([Fan et al. 2020](#))
- NLLB ([Costa-jussà et al. 2022](#))
- Google Translate

$\Delta$  from fine-tune XLM-R on English baseline



en → Lang.	Fine-tune <sub>en</sub>		NLLB+Word Align			NLLB+Markers		GMT+Word Align			GMT+Markers		
	Ref.	XLM <sub>R</sub>	QAali.	Awes.	Awesft	XML	EProj. ( $\Delta_{XLM_R}$ )	QAali.	Awes.	Awesft	XML	EProj. ( $\Delta_{XLM_R}$ )	
NER	yo	41.3	37.1	-	73.2	78.0	68.7	77.7 (+40.6)	-	72.1	66.1	71.8	73.8 (+36.7)
	ja	18.3	18.0	19.3	23.4	22.4	17.3	45.5 (+27.5)	19.3	23.0	22.6	42.0	43.5 (+25.5)
	zh	25.8	27.1	47.6	36.0	34.0	46.2	46.6 (+19.5)	45.2	43.3	39.6	43.8	45.9 (+18.8)
	th	1.5	0.7	-	2.6	2.5	8.8	14.0 (+13.3)	-	1.2	1.3	14.7	15.1 (+14.4)
QA	ur	54.2	63.6	-	71.6	71.8	74.4	74.7 (+11.1)	-	70.2	72.3	76.3	74.7 (+11.1)
	he	54.1	56.0	-	58.3	58.1	61.1	63.4 (+7.4)	-	59.6	60.2	63.7	67.1 (+11.1)
	ms	69.8	64.1	-	69.4	72.7	74.6	73.9 (+9.8)	-	73.0	73.8	73.2	74.1 (+10.0)
	my	51.3	53.5	-	61.6	62.9	56.2	60.1 (+6.6)	-	60.2	60.1	57.0	62.0 (+8.5)
	ar	43.7	48.5	49.3	48.7	47.6	45.9	50.5 (+2.0)	50.7	50.9	51.2	51.3	56.3 (+7.8)
	jv	58.4	62.3	-	64.8	61.6	67.7	67.0 (+4.7)	-	64.6	68.8	69.2	69.8 (+7.5)
	tl	72.2	73.0	-	80.1	78.8	79.5	79.3 (+6.3)	-	80.4	80.4	79.9	80.0 (+7.0)
	hi	71.0	69.5	-	73.9	73.4	73.8	74.4 (+4.9)	-	75.6	76.0	75.9	75.7 (+6.2)
	ka	68.9	68.8	-	74.5	75.0	70.4	74.2 (+5.4)	-	73.5	73.2	72.7	74.7 (+5.9)
	bn	76.3	75.1	-	80.5	80.2	80.1	80.7 (+5.6)	-	82.0	81.7	80.6	80.9 (+5.8)
	ta	56.9	58.8	-	63.1	63.7	53.8	63.5 (+4.7)	-	62.4	63.2	63.9	64.3 (+5.5)
	eu	62.1	63.6	-	70.3	70.0	64.7	68.7 (+5.1)	-	69.8	66.5	67.5	69.0 (+5.4)
	ko	58.0	57.9	-	61.1	60.6	59.4	58.0 (+0.1)	-	62.9	62.4	61.7	61.9 (+4.0)
	mr	64.1	63.9	-	63.6	64.0	62.9	64.9 (+1.0)	-	62.6	61.2	64.0	67.1 (+3.2)
	sw	70.0	68.5	-	70.6	71.5	70.1	69.7 (+1.2)	-	70.2	71.5	72.2	70.7 (+2.2)
	vi	77.2	74.2	-	70.4	65.8	77.8	77.5 (+3.3)	-	70.4	67.2	77.5	76.0 (+1.8)
	te	52.3	55.6	-	57.7	56.3	51.8	55.9 (+0.3)	-	57.4	56.8	57.6	57.4 (+1.8)
	id	52.3	52.4	-	53.5	55.3	52.7	53.1 (+0.7)	-	52.7	55.0	57.3	53.9 (+1.5)
	ml	65.8	63.5	-	63.2	64.8	56.5	61.3 (-2.2)	-	61.9	63.0	68.1	64.3 (+0.8)
	es	68.8	74.8	-	72.2	70.2	73.3	71.7 (-3.1)	-	71.3	72.6	73.5	75.6 (+0.8)
	de	77.9	79.4	79.7	79.5	79.6	81.5	80.0 (+0.6)	79.5	80.0	79.4	79.8	80.2 (+0.8)
	kk	49.8	53.5	-	53.5	53.9	40.4	54.0 (+0.5)	-	53.2	55.1	51.3	54.2 (+0.7)
	fr	79.0	80.1	80.7	79.8	80.9	80.9	81.5 (+1.4)	79.6	80.7	79.4	81.5	80.8 (+0.7)
	af	77.6	78.6	-	79.3	78.4	79.1	79.4 (+0.8)	-	79.1	78.9	79.0	79.2 (+0.6)
	et	78.0	79.6	-	80.7	79.2	80.2	79.9 (+0.3)	-	80.2	79.6	78.6	80.1 (+0.5)
	hu	79.3	81.0	-	80.3	79.8	79.7	80.4 (-0.6)	-	79.9	79.7	80.6	80.7 (-0.3)
	fi	78.6	80.6	-	81.0	80.9	80.4	79.8 (-0.8)	-	80.7	79.7	78.8	80.3 (-0.3)
	it	81.1	81.3	-	80.5	80.5	81.9	81.2 (-0.1)	-	80.3	80.4	81.1	80.9 (-0.4)
	tr	78.9	80.3	-	80.6	81.0	80.1	79.5 (-0.8)	-	80.1	80.2	81.5	79.6 (-0.7)
	nl	84.3	84.1	-	83.4	83.3	83.0	83.4 (-0.7)	-	83.5	82.9	83.0	83.1 (-1.0)
	bg	81.2	82.1	-	80.2	78.8	81.9	82.5 (+0.4)	-	80.9	79.7	82.5	80.6 (-1.5)
	pt	79.6	82.0	-	80.9	80.4	82.6	81.9 (-0.1)	-	79.0	80.2	80.6	80.1 (-1.9)
	ru	71.5	71.1	-	68.9	68.1	70.0	70.3 (-0.8)	-	67.4	66.8	67.4	68.2 (-2.9)
	el	77.2	79.3	-	76.3	75.7	77.7	74.1 (-5.2)	-	73.1	75.2	76.2	75.0 (-4.3)
	fa	61.1	64.3	-	41.5	47.3	51.3	52.1 (-12.2)	-	52.9	52.4	45.5	52.0 (-12.3)
	AVG	63.3	64.3	-	66.4	66.4	66.1	68.4 (+4.1)	-	66.7	66.6	68.3	68.9 (+4.6)
QA	ko	31.9	56.1	-	36.9	36.4	64.8	67.7 (+11.6)	-	37.6	37.1	60.9	65.0 (+8.9)
	bn	64.0	66.0	-	71.1	72.6	63.7	69.6 (+3.6)	-	73.6	69.3	74.4	71.0 (+5.0)
	fi	70.5	69.7	-	74.9	74.0	73.0	73.3 (+3.6)	-	74.9	74.9	73.1	74.0 (+4.3)
	te	70.1	72.9	-	74.9	74.6	69.9	78.3 (+5.4)	-	75.9	69.9	77.0	77.0 (+4.1)
	ar	67.6	72.4	74.2	76.8	76.4	72.7	75.9 (+3.5)	74.0	76.3	76.6	75.8	76.4 (+4.0)
	sw	66.1	69.9	-	73.0	74.7	72.4	73.4 (+3.5)	-	72.3	73.4	73.6	73.5 (+3.6)
	ru	67.0	66.5	-	70.9	71.5	69.1	70.4 (+3.9)	-	71.6	69.7	70.2	69.8 (+3.3)
	id	77.4	78.0	-	81.6	81.1	79.6	80.3 (+2.3)	-	80.4	81.3	78.9	79.7 (+1.7)
	AVG	64.3	68.9	-	70.0	70.2	70.7	73.6 (+4.7)	-	70.3	69.0	73.0	73.3 (+4.4)

# **Backup slides for CODEC**

# A Lot More Experiments in the Paper

- Using different MT systems:  
NLLB (600m, 1.3b, 3b) , M2M, mBART50 many-to-many, Google Translate
- Using different encoder LLMs for Word Alignment, NER. Event Extraction:  
mBERT, mDebertaV3, AfroXLMR, Glot500 – specialized for African languages
- Compare to a modified version of beam search with the constrained search space
- And more ....

# Error Analysis

Underline marks the projection errors.



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

# Error Analysis

Underline marks the projection errors.



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

only marks sub-words  
as an entity

Augmented data in low-resource languages

# Error Analysis

Underline marks the projection errors.



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

only marks sub-words  
as an entity

Augmented data in low-resource languages

having difficulty  
to project multiple spans

# Experiment Results

CODEC outperforms EasyProject and Awes-align for event argument extraction on ACE2005.

**Results of different methods on ACE-2005 in the translate-train setting**

Lang.	FT <sub>en</sub>	Awes-align	EasyProject	CODEC
Arabic	44.8	48.3	45.4	<b>48.4</b>
Chinese	54.0	57.3	<b>59.7</b>	59.1
AVG	49.4	52.8	52.6	<b>53.8</b>

Awes-align: Alignment-based approach | EasyProject: marker-based approach

FT<sub>En</sub>: Fine-tune on English data only

- Event Extraction: mT5-large
- Translation: NLLB

# Two Label Projection Setups

CODEC can be used for both – this flexibility leads to even better performance!

- **Translate-train** (*EasyProject*, *Awes-align*, *CODEC*)

Translate from English dataset to augment the training data in the target language

- **Translate-test** (*Awes-align*, *CODEC*)

Faransi ni Angileteri dɔrɔn de ye  
Fischler ka lajini dəmə .

*Translate*

Only France and Britain  
backed Fischler 's proposal.

*Faransi<sub>LOC</sub>* ni *Angileteri<sub>LOC</sub>* dɔrɔn  
de ye *Fischler<sub>PER</sub>* ka lajini dəmə .

*Label Projection*

*English NER model*

Only *France<sub>LOC</sub>* and *Britain<sub>LOC</sub>*  
backed *Fischler<sub>PER</sub>* 's proposal.

# Experiment Results

CODEC outperforms EasyProject and Awesome-align for event argument extraction on ACE2005.

Average F1 over 18 African languages on the test set of MasakhaNER2.0

FT <sub>En</sub>	Translate-train			Translate-test	
	Awes-align	EasyProject	CODEC ( $\Delta_{FT}$ )	Awes-align	CODEC ( $\Delta_{FT}$ )
54.5	63.6	64.9	67.1 (+12.7)	65.8	<b>70.4 (+16.0)</b>

Awes-align: Alignment-based approach | EasyProject: marker-based approach

FT<sub>En</sub>: Fine-tune on English data only