
# Multilingual and Multicultural Capabilities of Large Language Models

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

Georgia Tech

**Google Research / Feb 11, 2025**

# NLP X Research Lab

Georgia Tech Machine Learning

## Generative AI

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

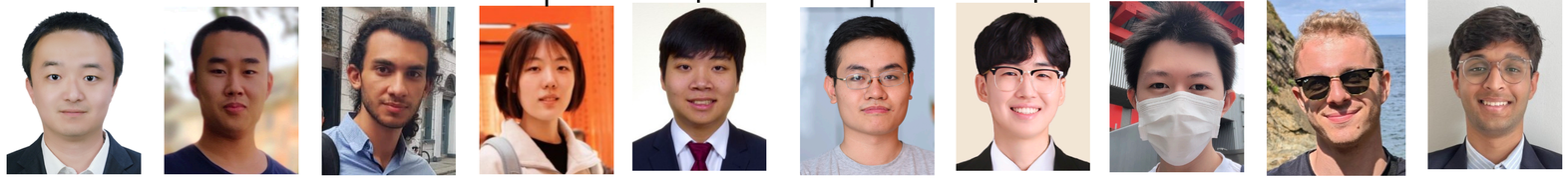
## Language Models

- multilingual multicultural adaptation
- decoding algorithms
- privacy, safety
- reasoning

## NLP+X Interdisciplinary Research

- HCI, human-centered NLP
- Education, Healthcare, Law, 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
Xiaofeng Wu, MS student
Romain Froger, MS student
Neel Kothari, MS student

Oleksandr Lavreniuk, Undergrad
Rachel Choi, Undergrad
Vishnesh Ramanathan, Undergrad
Govind Ramesh, Undergrad
Ian Ligon, Undergrad
Joseph Thomas, Undergrad
Nour Allah El Senary, Undergrad
Jad Bardawil, Undergrad
Tanish Patwa, MS student

# Today’s talk —

## 1. Constrained Decoding

**CODEC**


**(Le et al., ICLR 2024)**

Decoding algorithms for better performance on non-English languages.

## 2. Cultural Bias

**CAMEL**

**(Naous et al., ACL 2024)**

Measuring implicit cultural bias in LLMs and pre-training data.

## 3. Neologism

**NeoBench**

**(Zheng et al., ACL 2024)**

How well LLMs can handle newly emerging words?

# Today’s talk —

## 1. Constrained Decoding

**CODEC**


**(Le et al., ICLR 2024)**

Decoding algorithms for better performance on non-English languages.

## 2. Cultural Bias

**CAMEL**



**(Naous et al., ACL 2024)**

Measuring implicit cultural bias in LLMs and pre-training data.

## 3. Neologism

**NeoBench**



**(Zheng et al., ACL 2024)**

How well LLMs can handle newly emerging words?

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

**Masakhane NER 2.0**


| Lang. | GPT-4† | $FT_{En}$ |
|---|---|---|
| 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 | **55.2** | 35.8 |
| Mossi | 44.2 | 45.0 |
| Chichewa | 75.8 | **79.5** |
| chiShona | 66.8 | 35.2 |
| Kiswahili | 82.6 | **87.7** |
| Setswana | 62.0 | 64.8 |
| Akan/Twi | 52.9 | 50.1 |
| Wolof | 62.6 | 44.2 |
| isiXhosa | 69.5 | 24.0 |
| Yoruba | **58.2** | 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† | $FT_{En}$ | 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 | **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) |

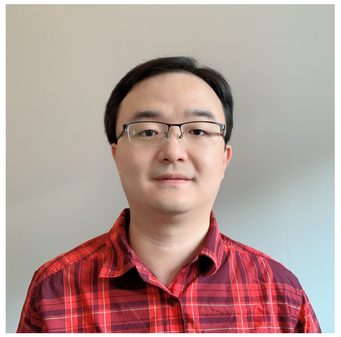
Label Projection methods

[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

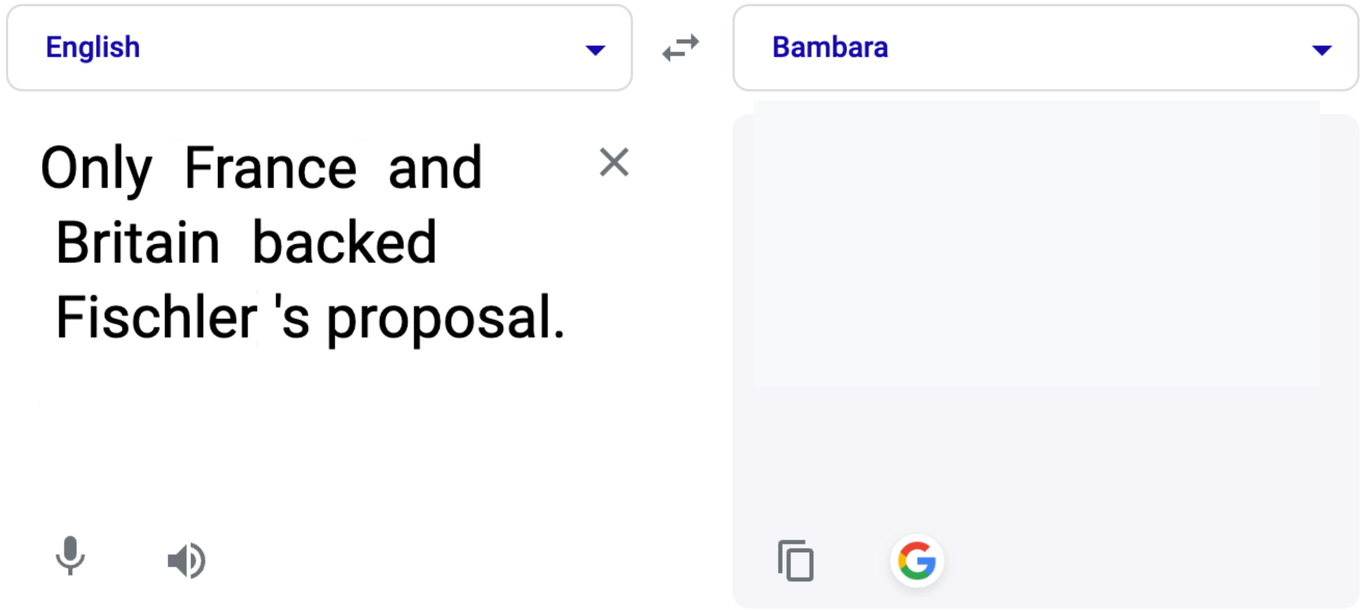
Alan Ritter

Wei Xu


A systematic study of marker-based approach for label projection

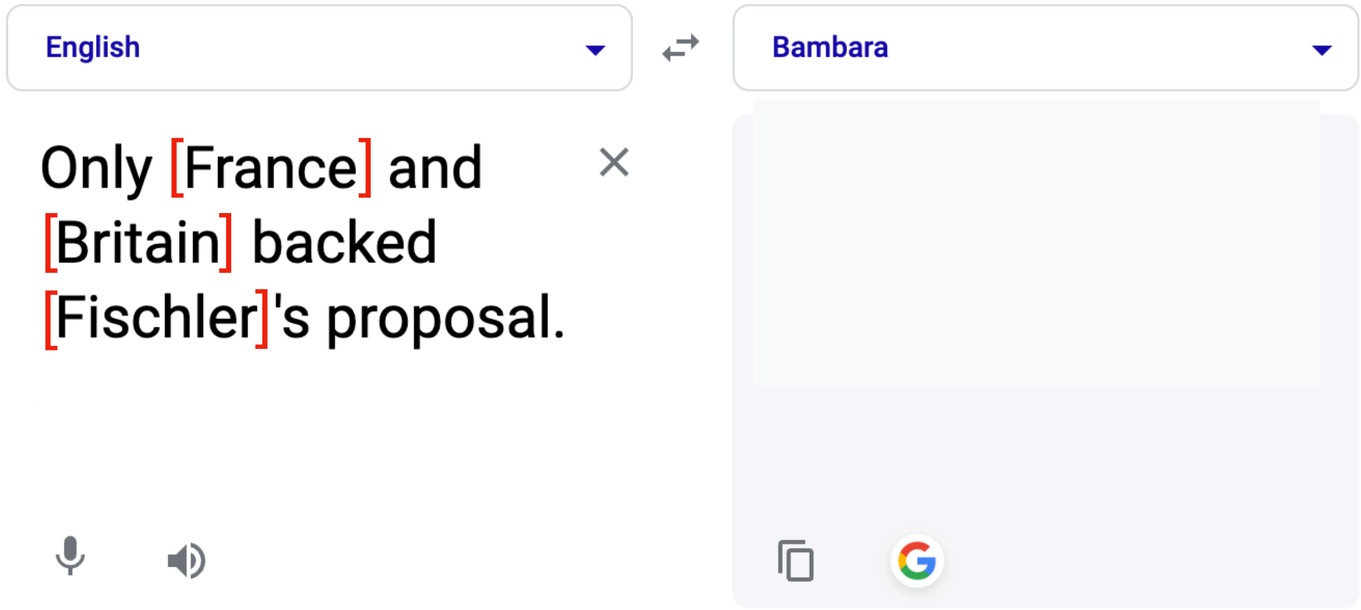
# Marker-based Approach

Translating annotated training data from one language to the other



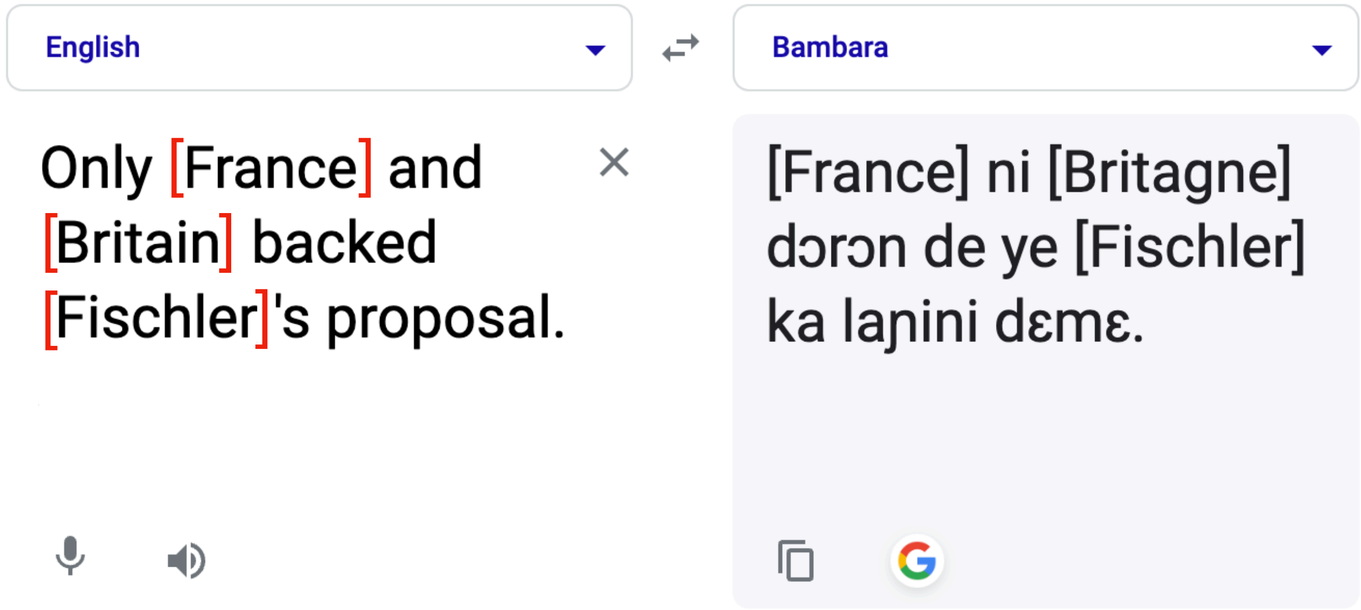
# Marker-based Approach

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



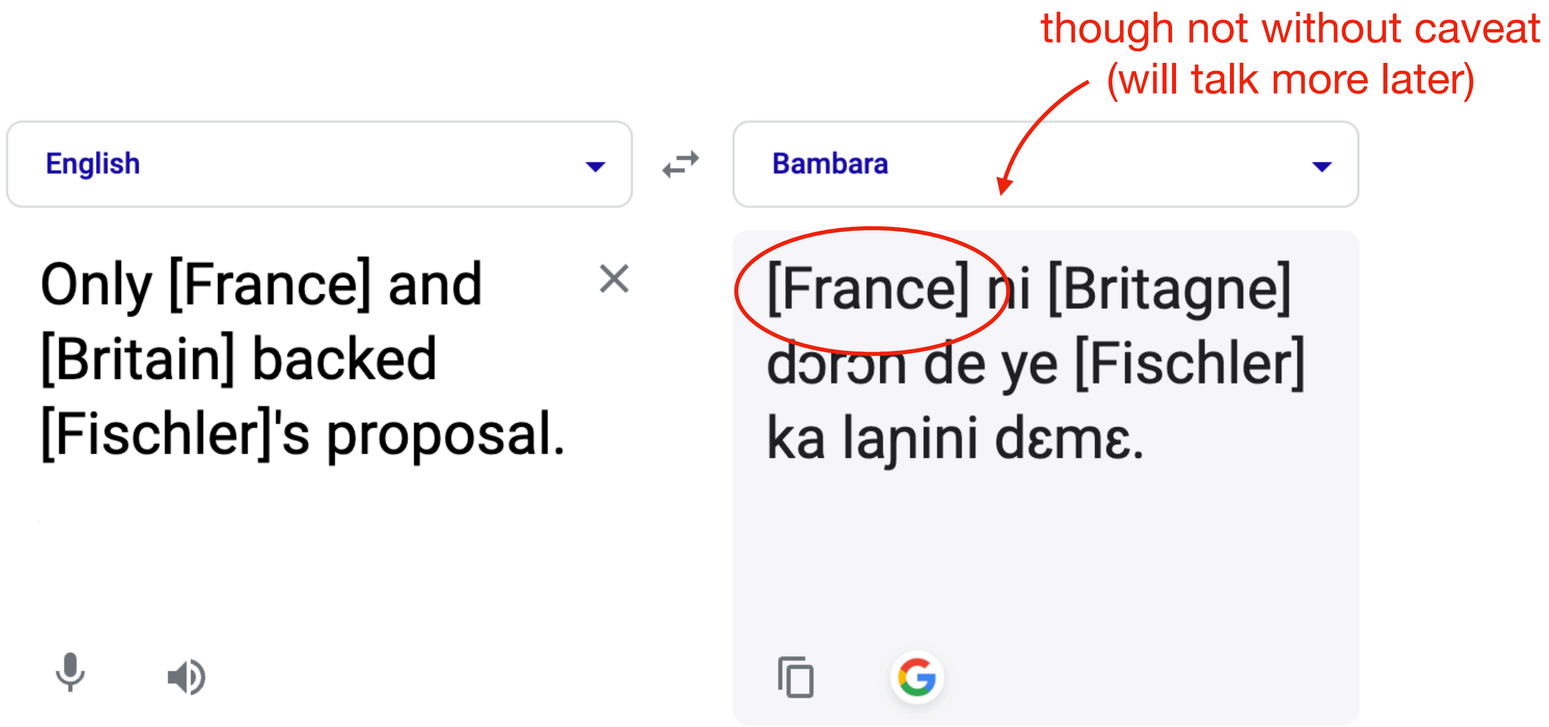
# Marker-based Approach

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.



# 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

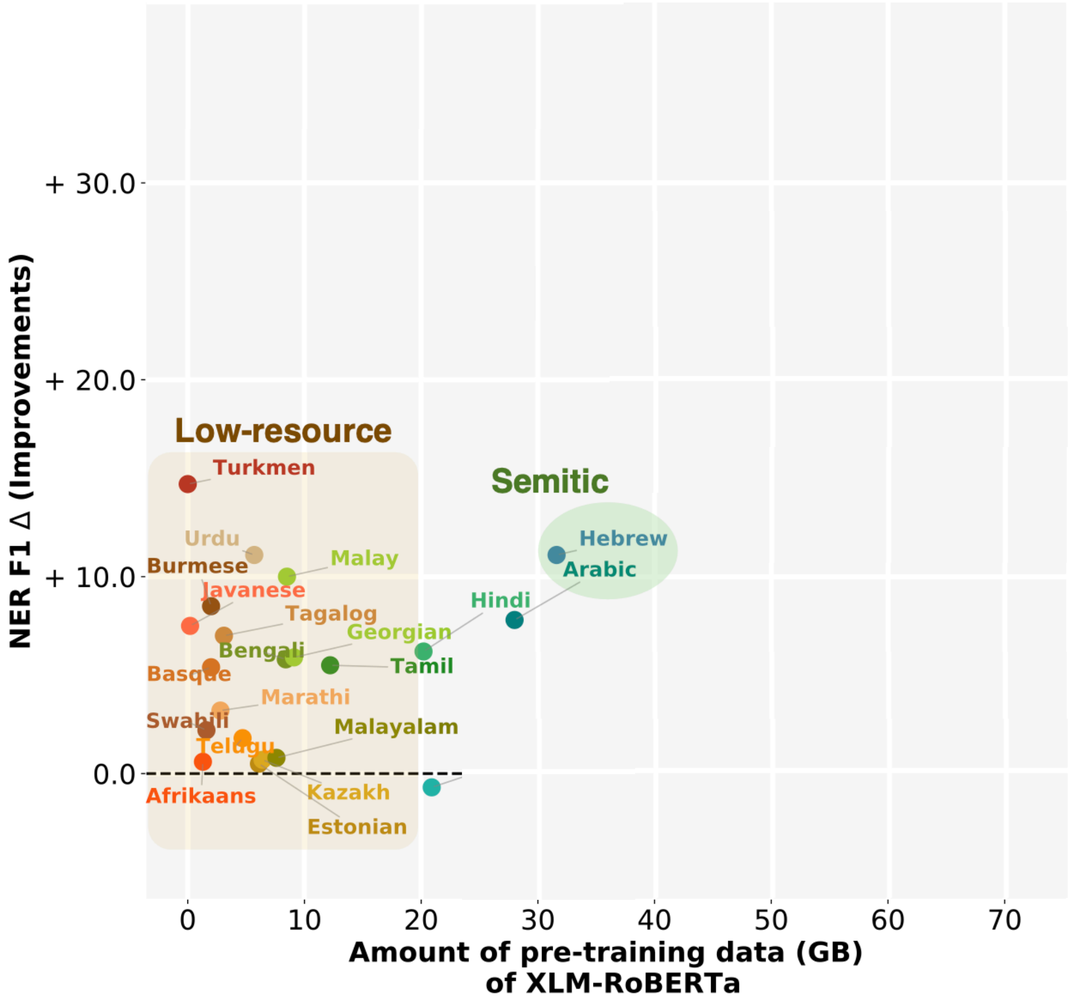
XML tags (e.g., <loc> </loc> )  or  [ ]  “ ”  ( )  < >  { }

- 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

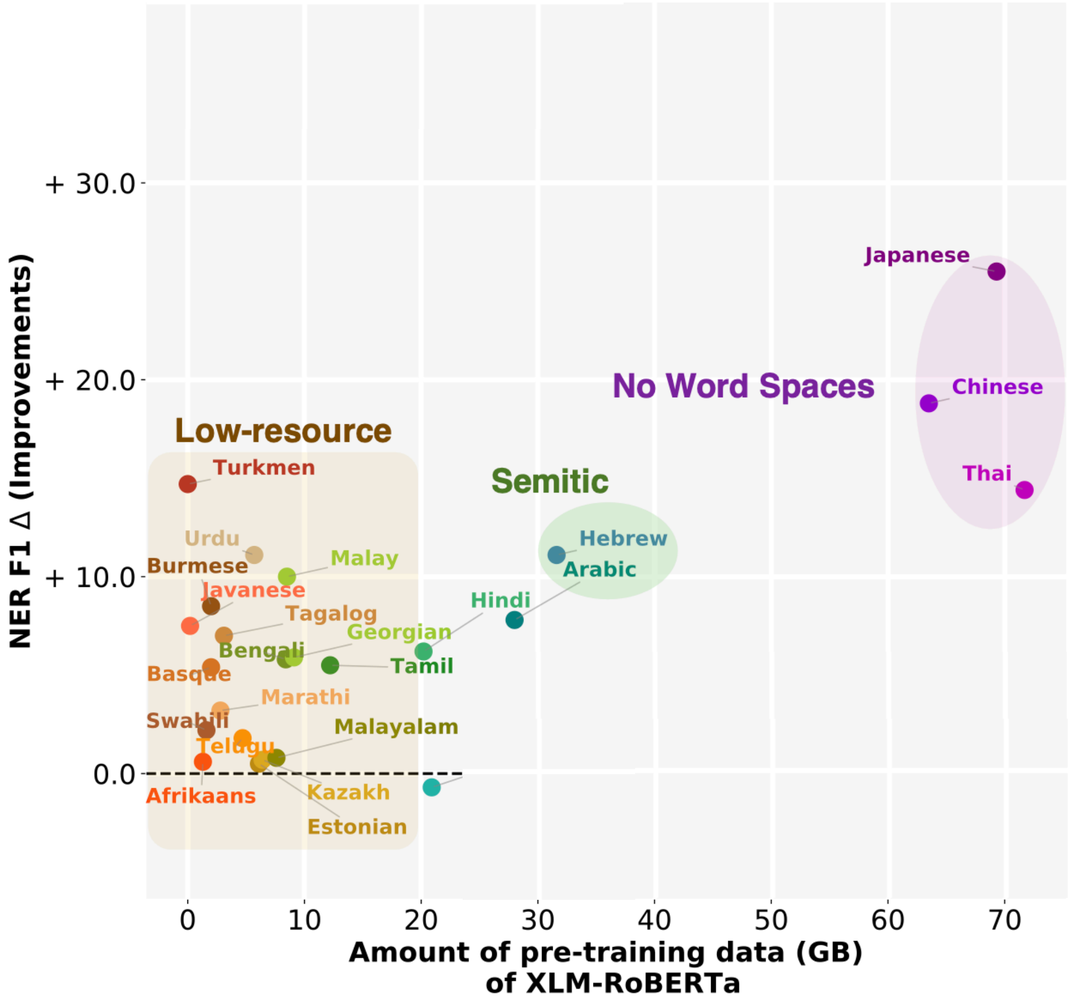
# EasyProject - Easy Marker-based Projection

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



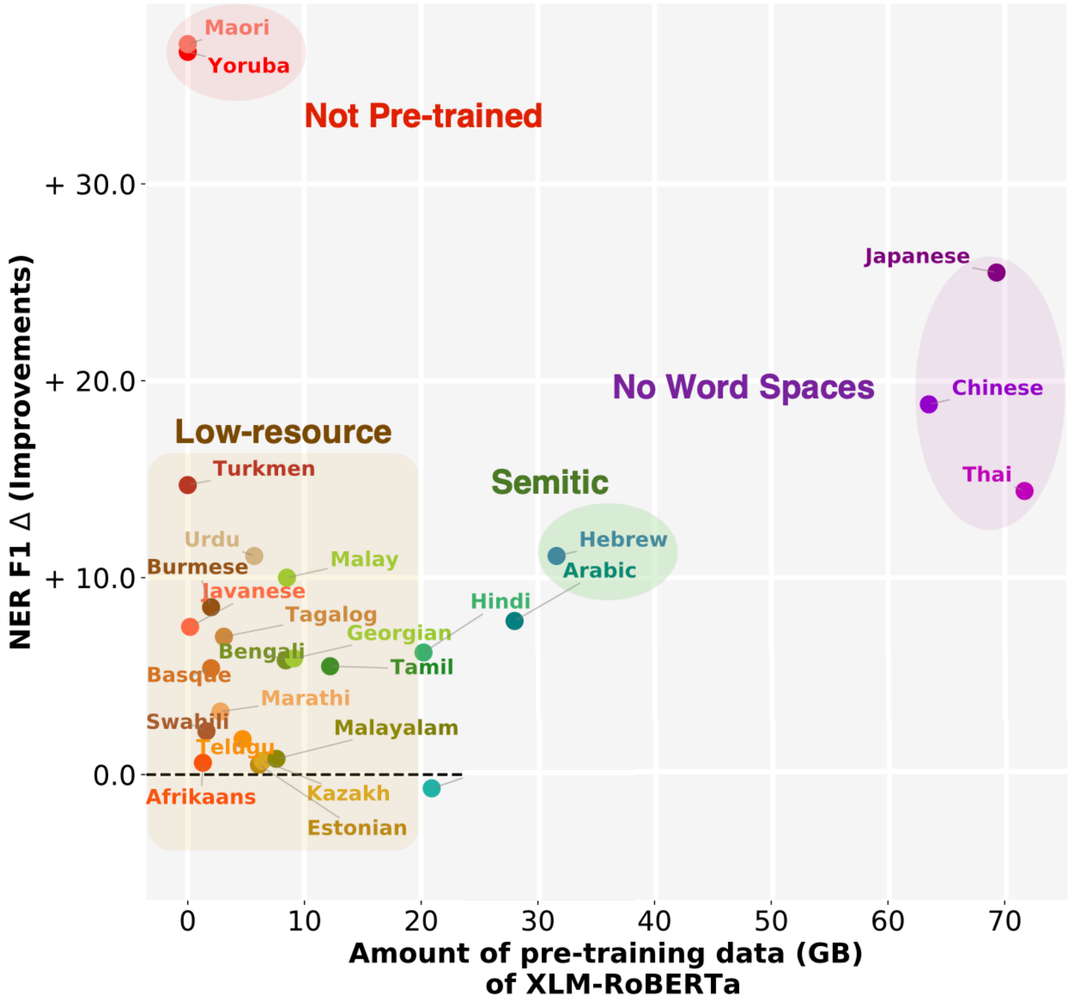
# EasyProject - Easy Marker-based Projection

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



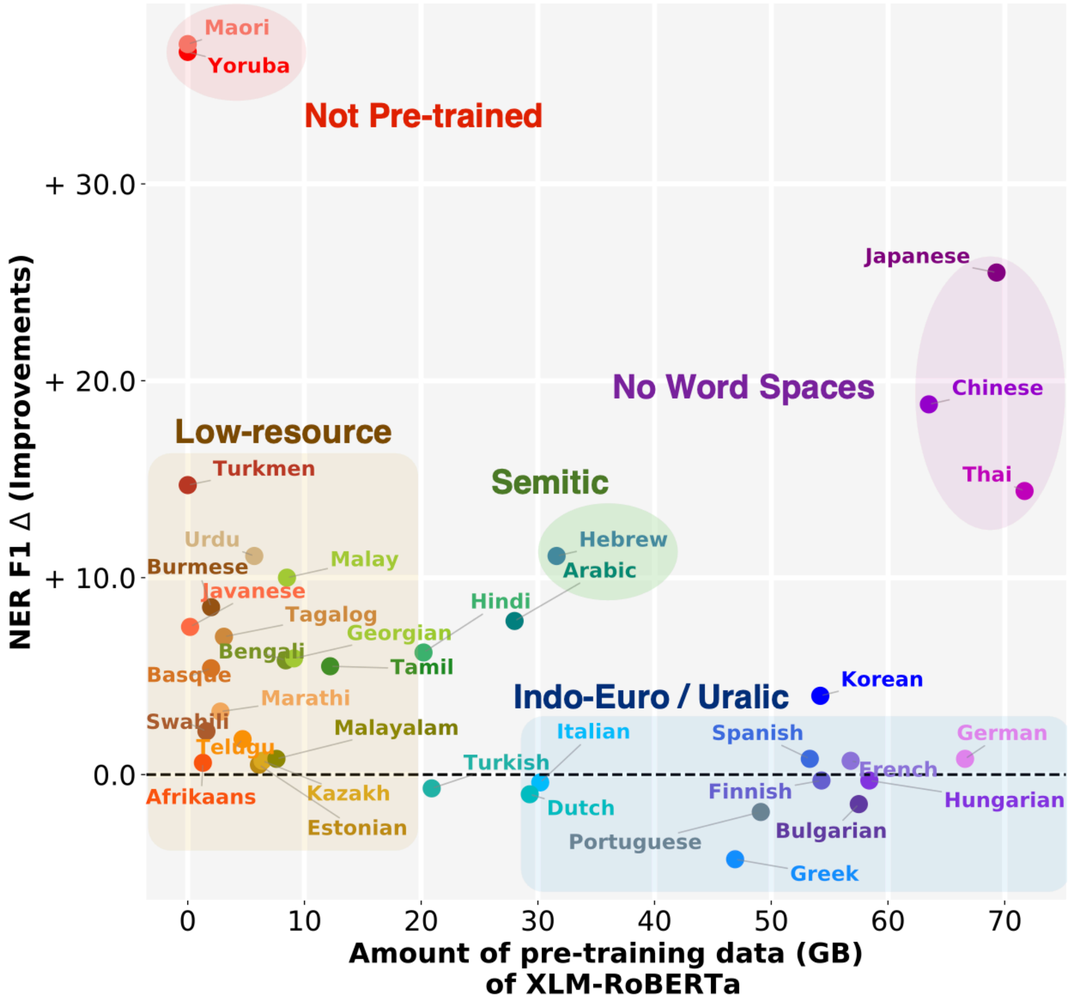
# EasyProject - Easy Marker-based Projection

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



# EasyProject - Easy Marker-based Projection

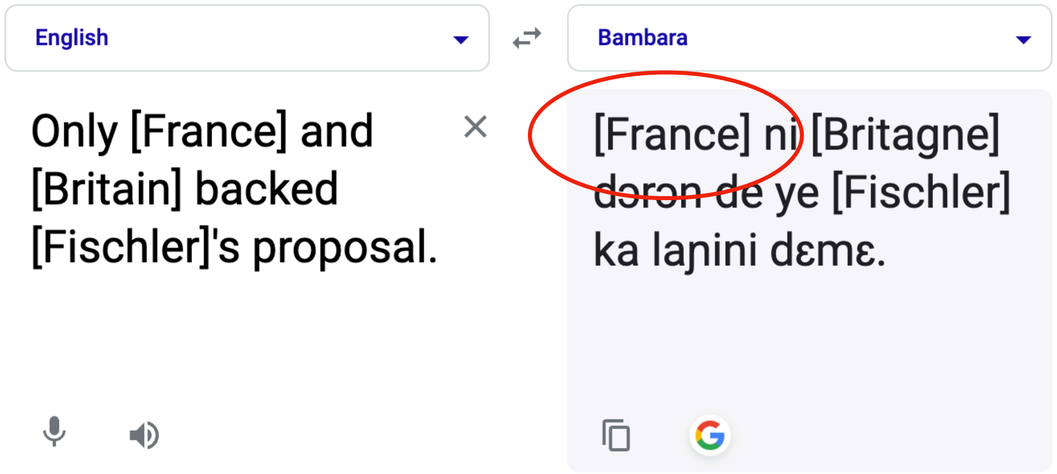
Especially promising for low-resource languages & languages that are written in non-Latin scripts



# Zero-shot Cross-lingual Label Projection

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

**marker-based approach**



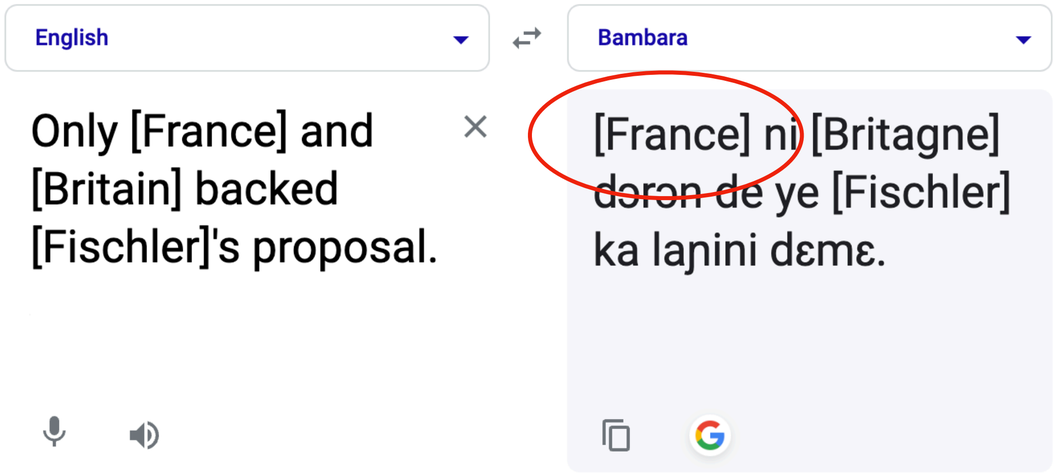

Only need a MT system & work surprisingly well !

But, degraded MT quality due to injected markers

# Zero-shot Cross-lingual Label Projection

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

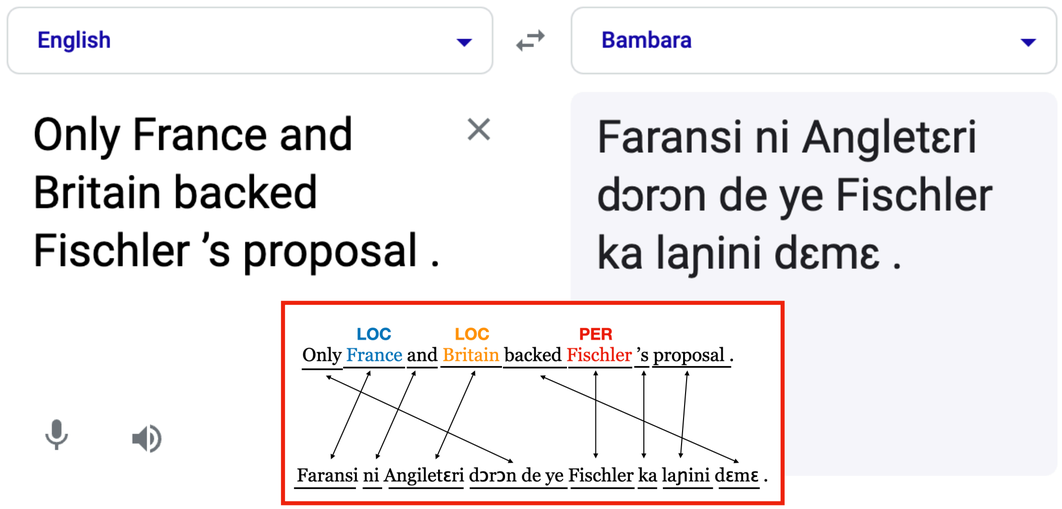
## marker-based approach




Only need a MT system & work surprisingly well !

But, degraded MT quality due to injected markers

## word alignment-based approach

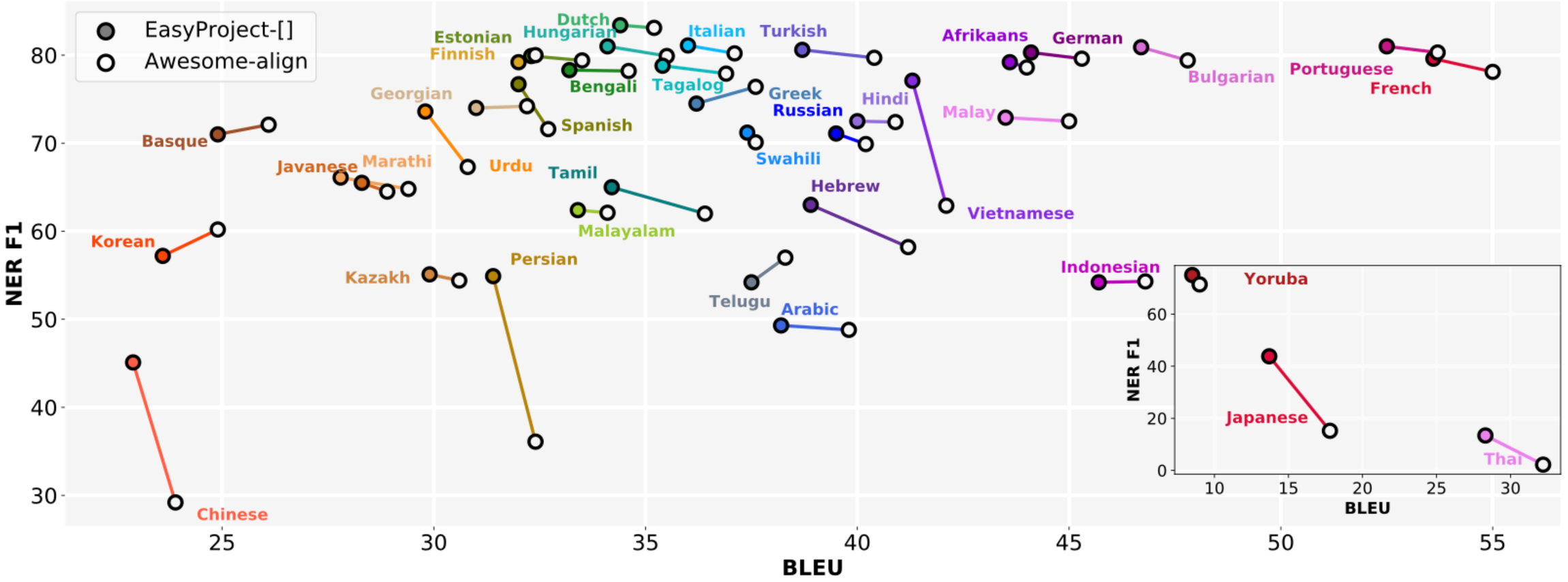



normally better MT quality

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 without scarifying the translation quality?**

# Constrained Decoding for Cross-lingual Label Projection (CODEC)

Duong Minh Le

Yang Chen

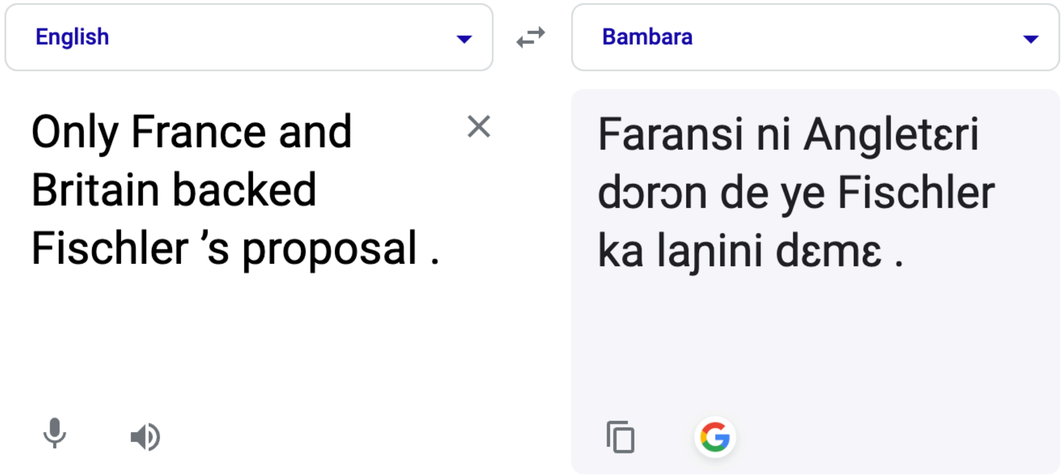
Alan Ritter

Wei Xu

A better technical solution for marker-based label projection

# Key Idea

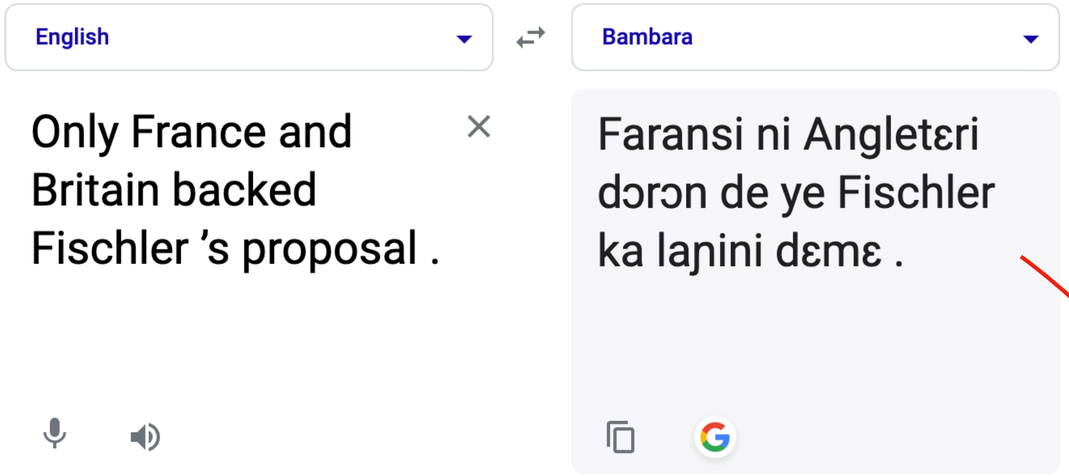
Step 1. Translate the original sentence as usual without markers.




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:**

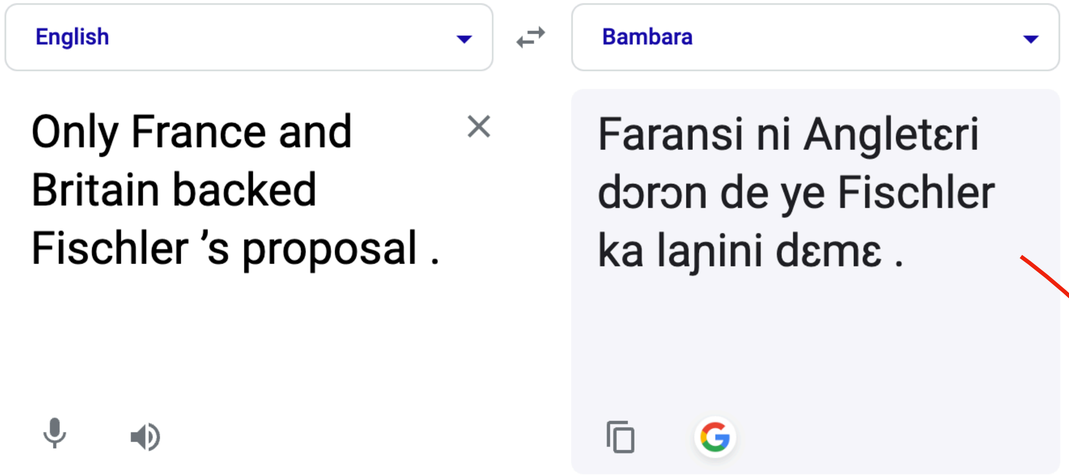
Only [France] and [Britain] backed [Fischler]'s proposal.

→ **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 $y^{tmpl}$ .

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

$O(n^{2m})$

# An Efficient Constrained Decoding Algorithm

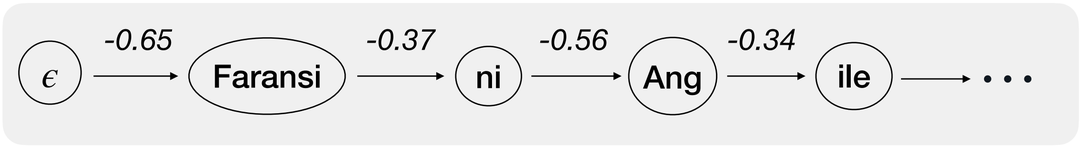
(1) Prune opening marker positions based on the contrastive log-likelihood difference.

# 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 .”

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



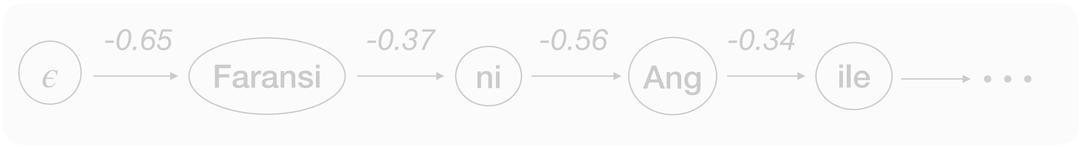

$p_2^i = \log P(y_i^{tmpl} | y_{<i}^{tmpl}, x^{mark})$ (Conditioned on source text w/ markers)

# An Efficient Constrained Decoding Algorithm

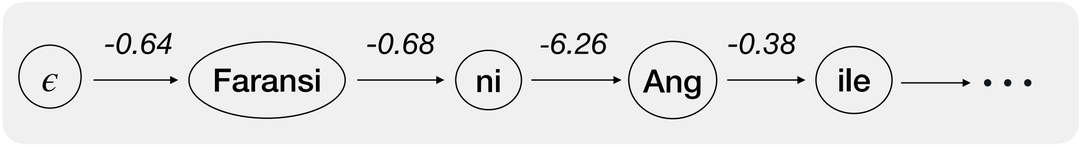
(1) Prune opening marker positions based on the contrastive log-likelihood difference.

**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 .”

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



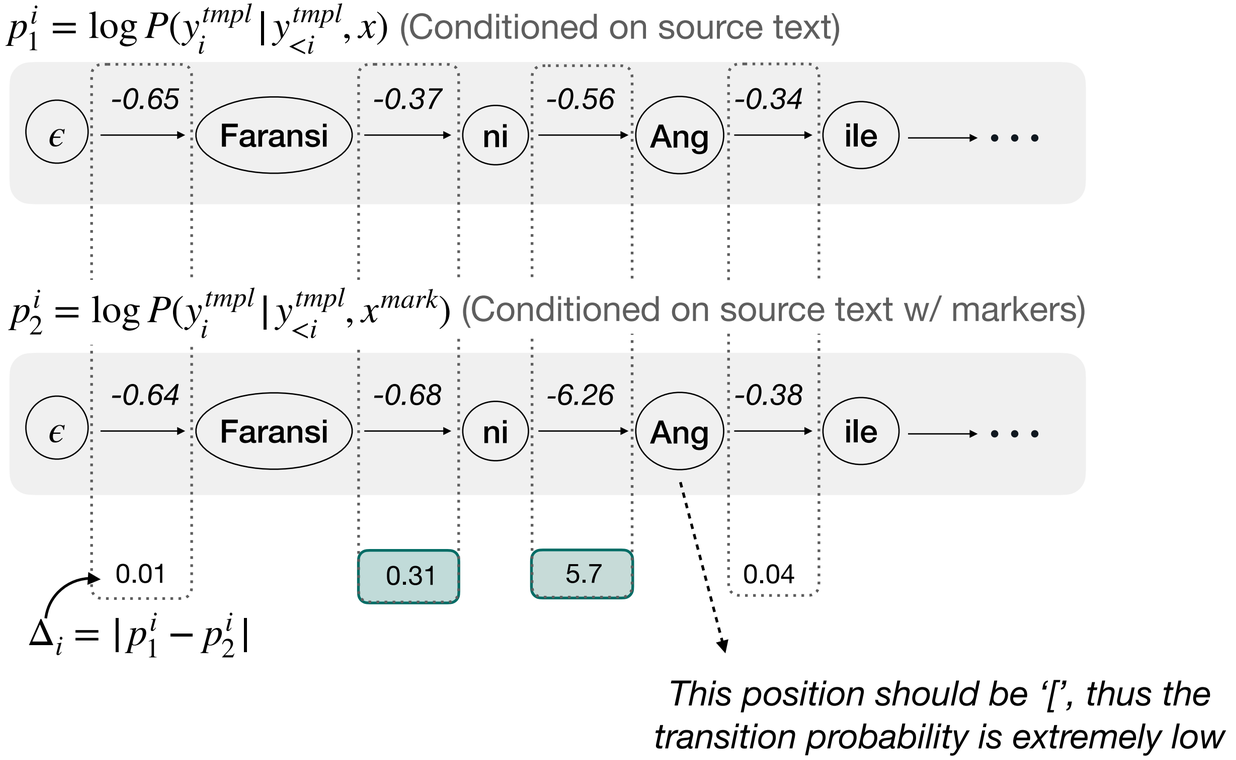

$p_2^i = \log P(y_i^{tmpl} \mid y_{<i}^{tmpl}, x^{mark})$ (Conditioned on source text w/ markers)



# An Efficient Constrained Decoding Algorithm

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

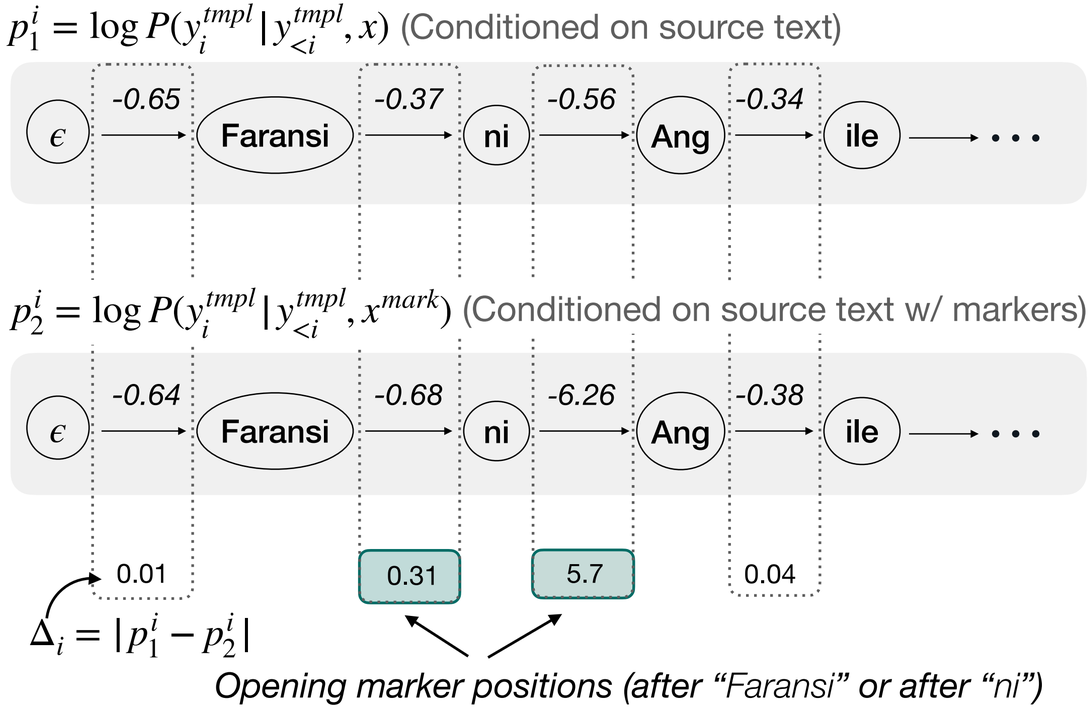
**Input:** $x$ = “Only France and Britain backed Fischler 's proposal .” $x^{mark}$ = “Only France and [ Britain ] backed Fischler 's proposal .” $y^{tmpl}$ = “Faransi ni Angiletɛri dɔrɔn de ye Fischler ka laɲini dɛmɛ .”



# 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 $L_d^k = \log P(y_{1:d}^k | x^{mark})$ .

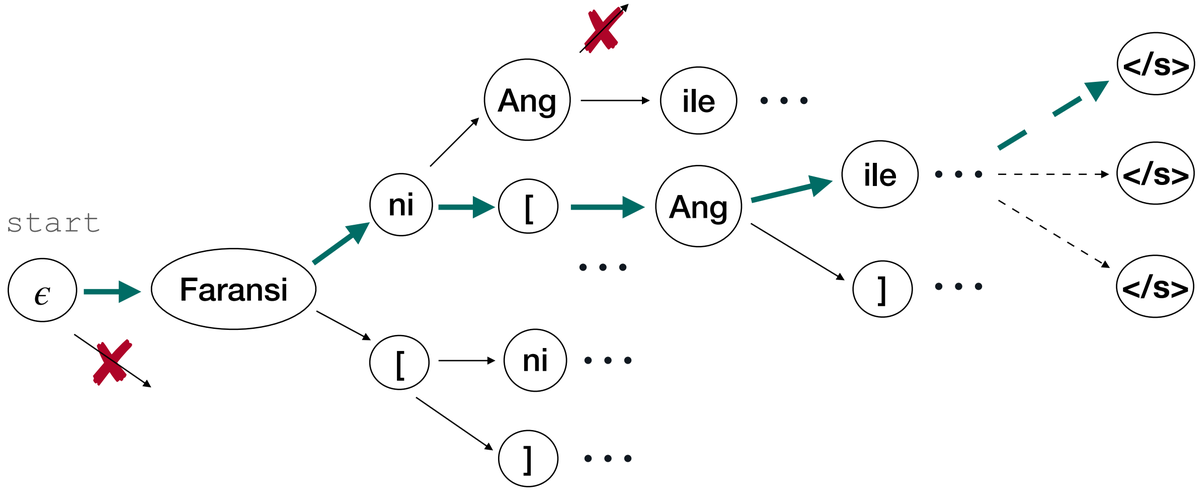
$$d = \min\left(\max\left(j + \delta, q\right), |y^k|\right)$$

# 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 Angiletɛri dɔrɔn de ye Fischler ka laɲini dɛmɛ .”



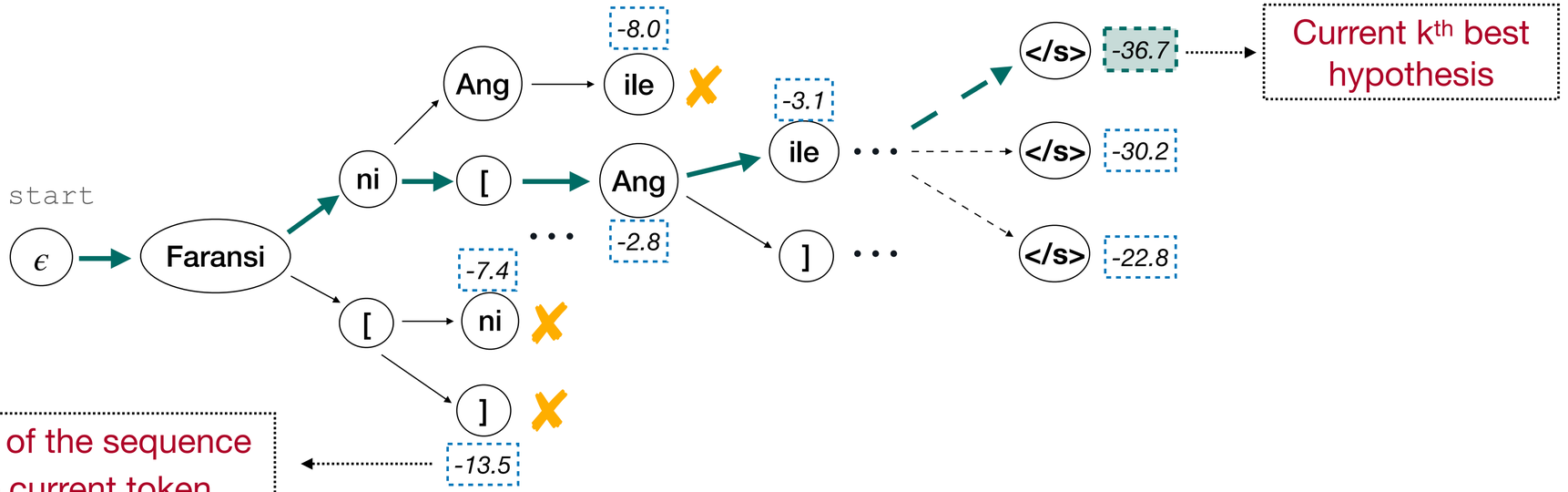

✘ *Prune opening-marker positions*

# 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 ."

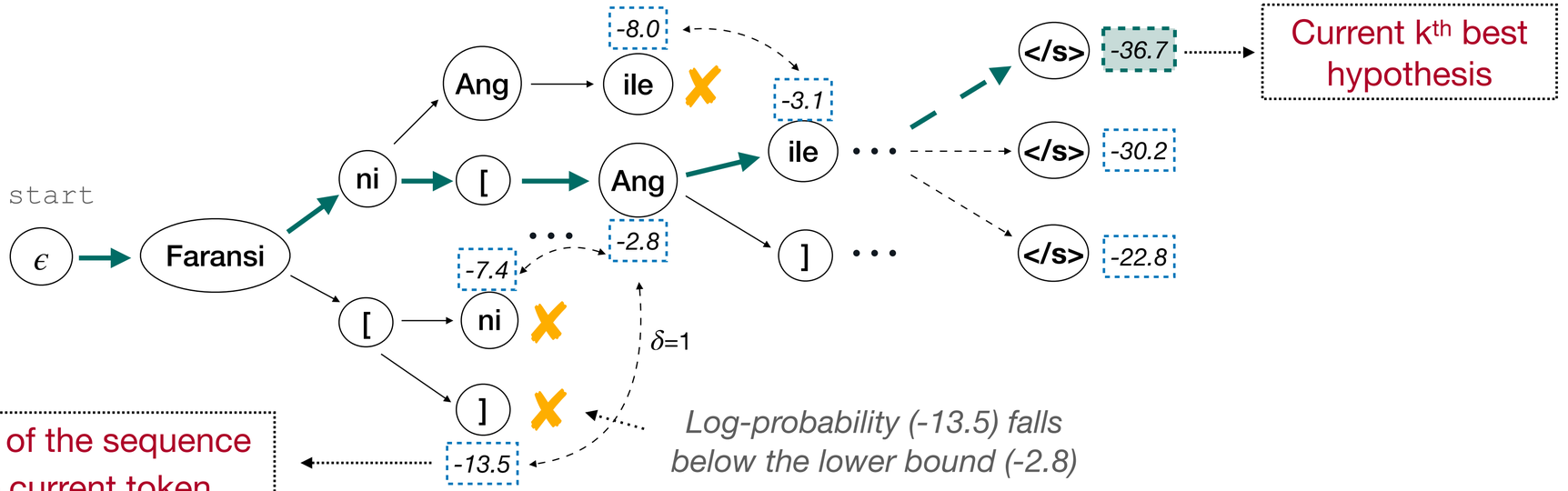


# An Efficient Constrained Decoding Algorithm

(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 Angiletɛri 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: $\epsilon$), $y^{tmpl}$: 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, $\delta$: lower bound hyperparameter

$flag \leftarrow$ {check if all markers are generated}
**if** $y_{|y|} =$ </s> and $flag =$ TRUE: **then**
    $H$.push($(L_{|y|}, L, y)$) ▷ $H$ sorts by the first element
    **if** len($H$) $> k$ **then**
        $H$.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 ($\Delta_{FT}$) |
| Bambara | 46.8 | 37.1 | 45.0 | 45.8 | 45.8 (+8.7) |
| Ewe | 75.5 | 75.3 | 78.3 | 78.5 | **79.1** (+3.8) |
| Fon | 19.4 | 49.6 | 59.3 | 61.4 | **65.5** (+15.9) |
| Hausa | 70.7 | 71.7 | 72.7 | 72.2 | 72.4 (+0.7) |
| Igbo | 51.7 | 59.3 | 63.5 | 65.6 | 70.9 (+11.6) |
| Kinyarwanda | 59.1 | 66.4 | 63.2 | 71.0 | 71.2 (+4.8) |
| Luganda | 73.7 | 75.3 | 77.7 | 76.7 | 77.2 (+1.9) |
| Luo | **55.2** | 35.8 | 46.5 | 50.2 | 49.6 (+13.8) |
| Mossi | 44.2 | 45.0 | 52.2 | 53.1 | **55.6** (+10.6) |
| Chichewa | 75.8 | **79.5** | 75.1 | 75.3 | 76.8 (-2.7) |
| chiShona | 66.8 | 35.2 | 69.5 | 55.9 | 72.4 (+37.2) |
| Kiswahili | 82.6 | **87.7** | 82.4 | 83.6 | 83.1 (-4.6) |
| Setswana | 62.0 | 64.8 | 73.8 | 74.0 | 74.7 (+9.9) |
| Akan/Twi | 52.9 | 50.1 | 62.7 | 65.3 | 64.6 (+14.5) |
| Wolof | 62.6 | 44.2 | 54.5 | 58.9 | 63.1 (+18.9) |
| isiXhosa | 69.5 | 24.0 | 61.7 | **71.1** | 70.4 (+46.4) |
| Yoruba | **58.2** | 36.0 | 38.1 | 36.8 | 41.4 (+5.4) |
| isiZulu | 60.2 | 43.9 | 68.9 | 73.0 | **74.8** (+30.9) |
| AVG | 60.4 | 54.5 | 63.6 | 64.9 | 67.1 (+12.7) |

- NER: mDeBERTa-v3
- MT: NLLB

prior marker-based approach cannot do this

# Experiment Results

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

| Lang. | GPT-4† | $FT_{En}$ | Translate-train | | | Translate-test | |
|---|---|---|---|---|---|---|---|
| | | | Awes-align | EasyProject | CODEC ($\Delta_{FT}$) | Awes-align | CODEC ($\Delta_{FT}$) |
| 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) |

# 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 ….

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

# 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.[1]

## 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 candidate 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-

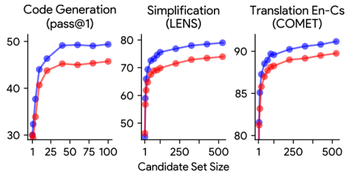
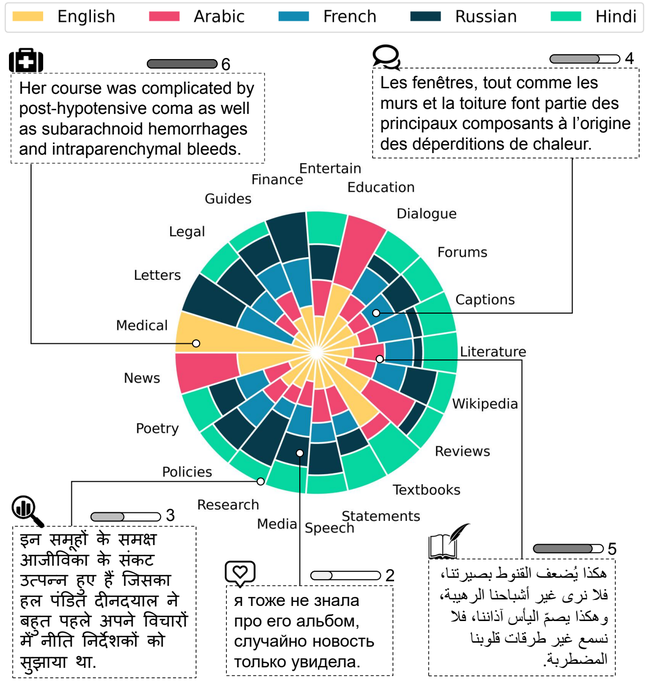



Figure 1: Multi-prompt and single prompt MBR results for code generation on HUMANEVAL, text simplification on SIMPEVAL, and translation on WMT '22 EN-CS generated with open-source 7B LLMs (details in §4).



### 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






### 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

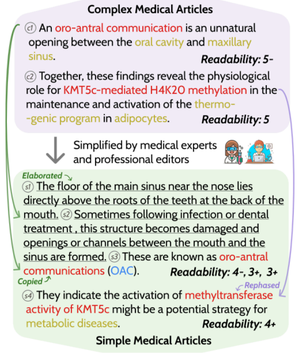



Figure 1: An illustration of our dataset, with sentence readability ratings and fine-grained complex span annotation on 4,520 sentences, including "Google-Hard" and "Google-Easy", abbreviations, and general complex terms, etc. We also analyze how medical jargon are being handled during simplification. 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. Constrained Decoding

**CODEC**


**(Le et al., ICLR 2024)**

Decoding algorithms for better performance on non-English languages.

## 2. Cultural Bias

**CAMEL**



**(Naous et al., ACL 2024)**

Measuring implicit cultural bias in LLMs and pre-training data.

## 3. Neologism

**NeoBench**



**(Zheng et al., ACL 2024)**

How well LLMs can handle newly emerging words?

# Today’s talk —

## 1. Constrained Decoding

CODEC


(Le et al., ICLR 2024)

Decoding algorithms for better performance on non-English languages.

## 2. Cultural Bias

**CAMEL**

**(Naous et al., ACL 2024)**

Measuring implicit cultural bias in LLMs and pre-training data.

## 3. Neologism

NeoBench



(Zheng et al., ACL 2024)

How well LLMs can handle newly emerging words?

# 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?**

| | Llama3.3-70b | | | | | | $\text{XLMR}_{\text{large}}$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Arabic | | | English | | | Arabic | | | English | | |
| | *Arab* | *Western* | ΔAcc | *Arab* | *Western* | ΔAcc | *Arab* | *Western* | ΔF1 | *Arab* | *Western* | ΔF1 |
| Authors | 92.62 | 90.28 | -2.34 | 98.99 | 99.16 | 0.17 | 86.80 | **87.93** | 1.13 | 95.64 | 94.98 | -0.66 |
| Beverage | 82.65 | 78.19 | -4.46 | 99.14 | 97.71 | -1.43 | 63.06 | **72.86** | 9.80 | 92.06 | 89.77 | -3.29 |
| Food | 84.08 | **84.71** | 0.63 | 95.84 | 98.21 | 2.37 | 63.76 | **73.59** | 9.83 | 91.57 | 90.45 | -1.12 |
| Location | 80.66 | **95.59** | 14.93 | 98.58 | 99.89 | 1.31 | 64.07 | **91.32** | 27.25 | 89.54 | 95.71 | 6.17 |
| Names (F) | 63.38 | **77.39** | 14.01 | 99.86 | 99.14 | -0.72 | 62.22 | **82.65** | 20.43 | 97.87 | 96.36 | -1.51 |
| Names (M) | 75.45 | **76.23** | 0.78 | 99.43 | 99.78 | 0.35 | 80.09 | **85.03** | 4.94 | 94.01 | 93.13 | -0.88 |
| Sports | 68.58 | **79.01** | 10.43 | 92.77 | 96.02 | 3.25 | 74.52 | **84.14** | 9.62 | 92.14 | 93.12 | 0.97 |
| Religious | 82.49 | **82.98** | 0.49 | 98.52 | 97.69 | -0.83 | 95.30 | **97.13** | 1.83 | 94.34 | 95.76 | 1.42 |

Table 3: Average performance of Llama3.3-70b (QA Accuracy ↑) and $\text{XLMR}_{\text{large}}$ (NER F1 ↑) on Arab and Western entities when tested in Arabic and English. More results with Aya23-35b and AceGPTv2-70b are in Appendix. ΔAcc and ΔF1 are 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.

Tarek Naous, Wei Xu. “On The Origin of Cultural Biases in Language Models: From Pre-training Data to Linguistic Phenomena” (NAACL 2025)

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

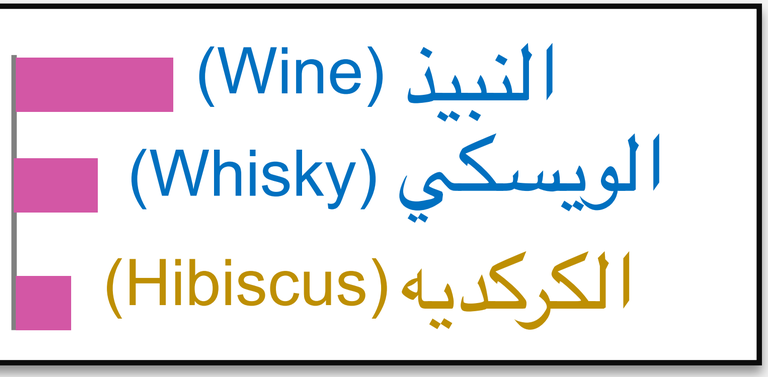
E.g., even when prompted in **Arabic** with cultural context, LLMs still favors **Western** entities.

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

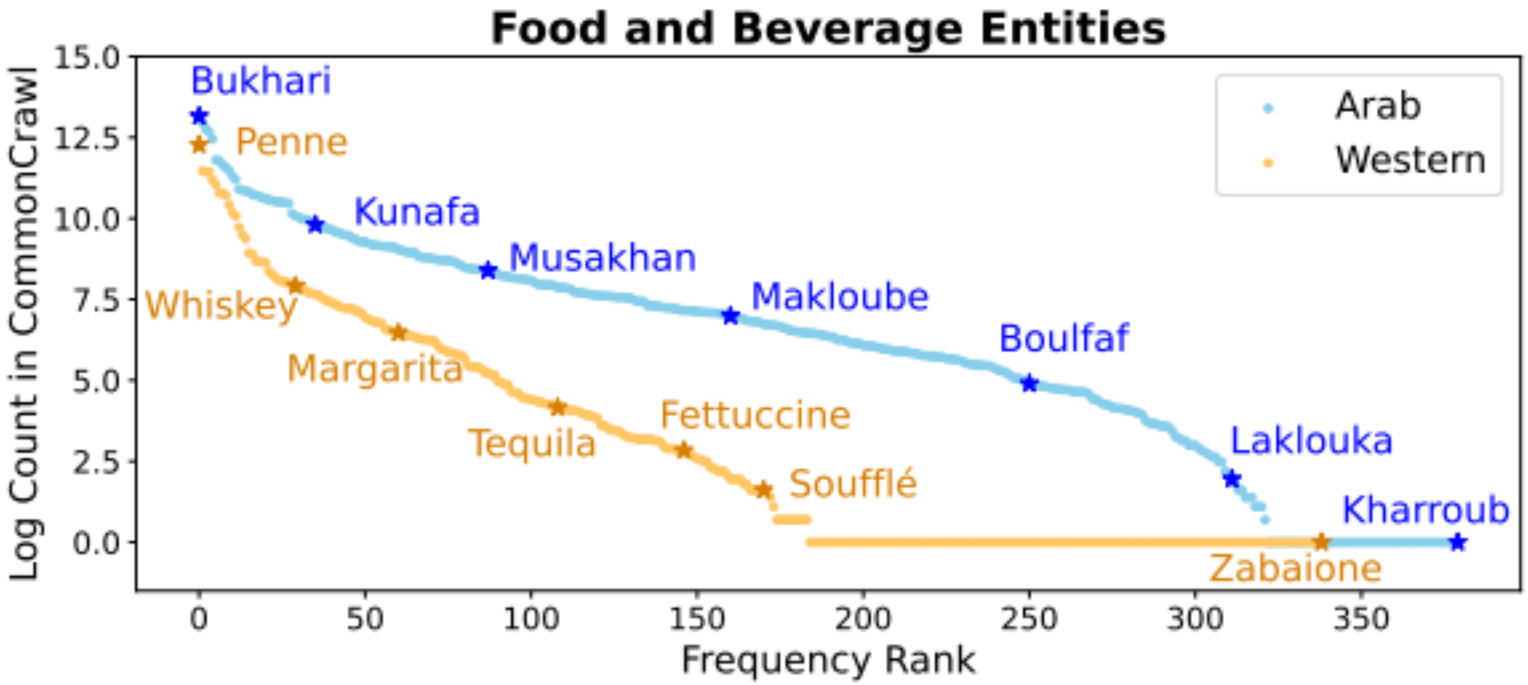
20k cultural relevant entities spanning 8 categories that contrast **Arab** vs. **Western** cultures.

| | |
|---|---|
| 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]<br>(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<br>(I ate [MASK] and it's worse than anything you can ever have) |
| كنت اصلي القيام في [MASK] و القارئ تلاوته للقرآن تأسر القلب<br>(I was praying Qiyam in [MASK] and the Quraan recitation captivated my heart) | كان معزوم في حفل زفاف شاب في [MASK]<br>(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]*

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

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

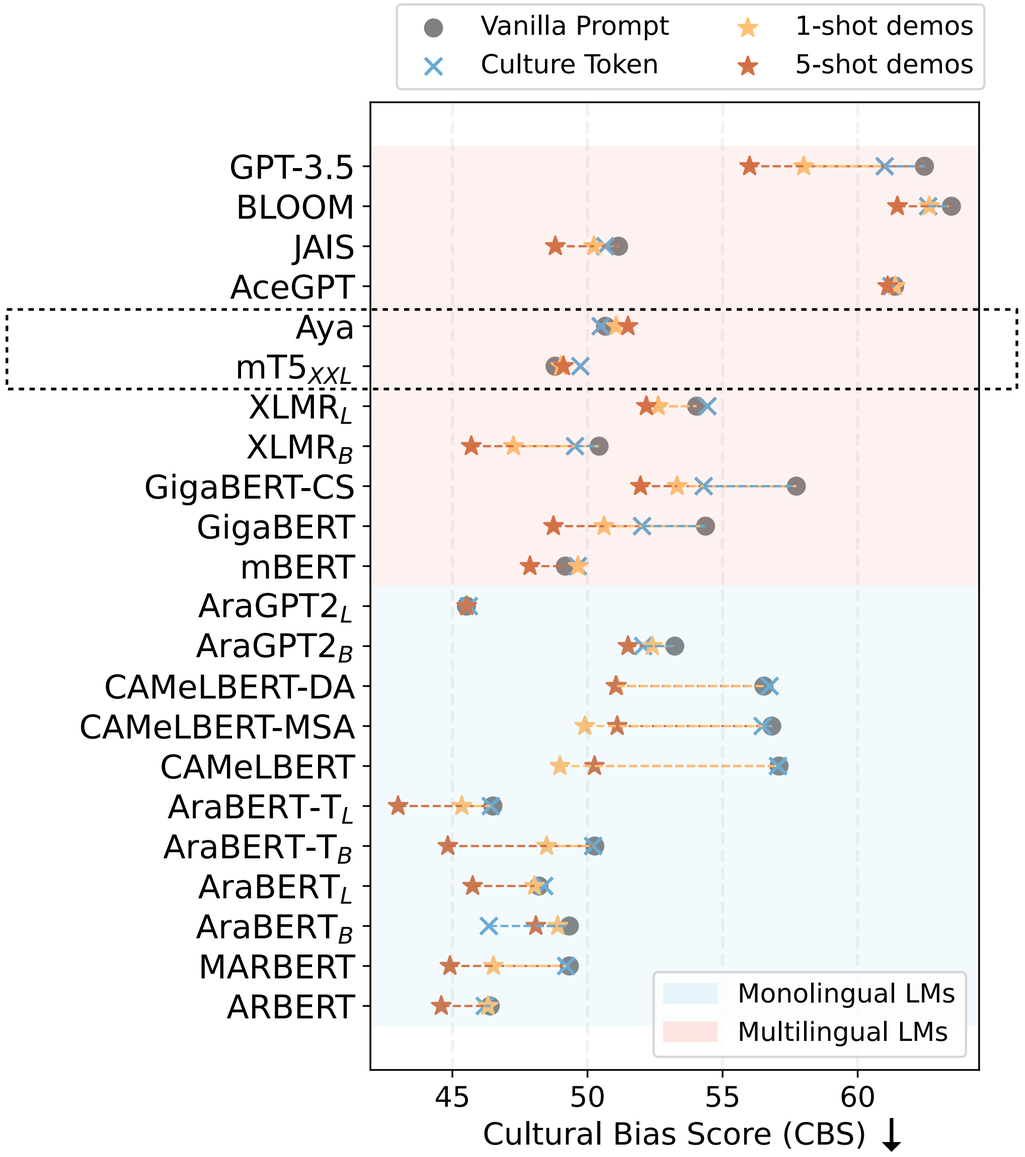
Prompt Set $T = \{t_k\}_{k=1}^{K}$

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

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

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

# CAMeL — How often LLMs favor Western entities?




A set of prompts $T = \{t_k\}_{k=1}^{K}$,

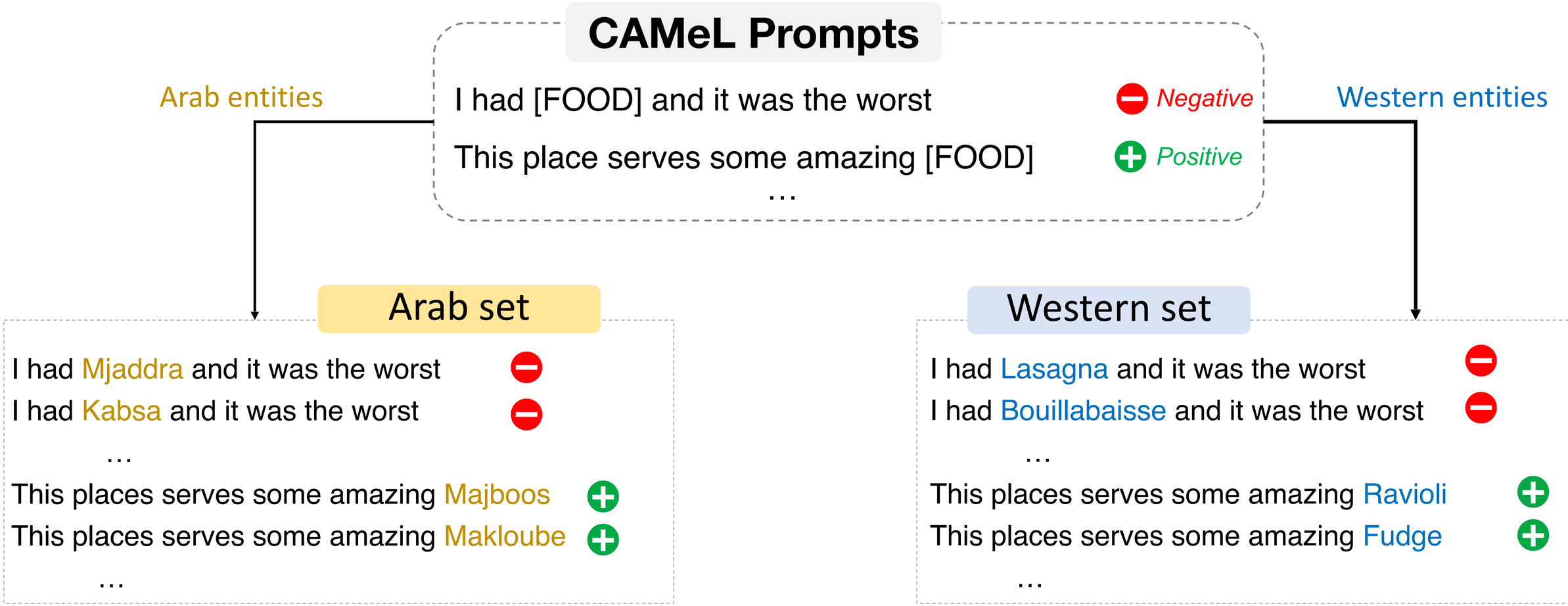
Arab entities $A = \{a_i\}_{i=1}^{N}$ and

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

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

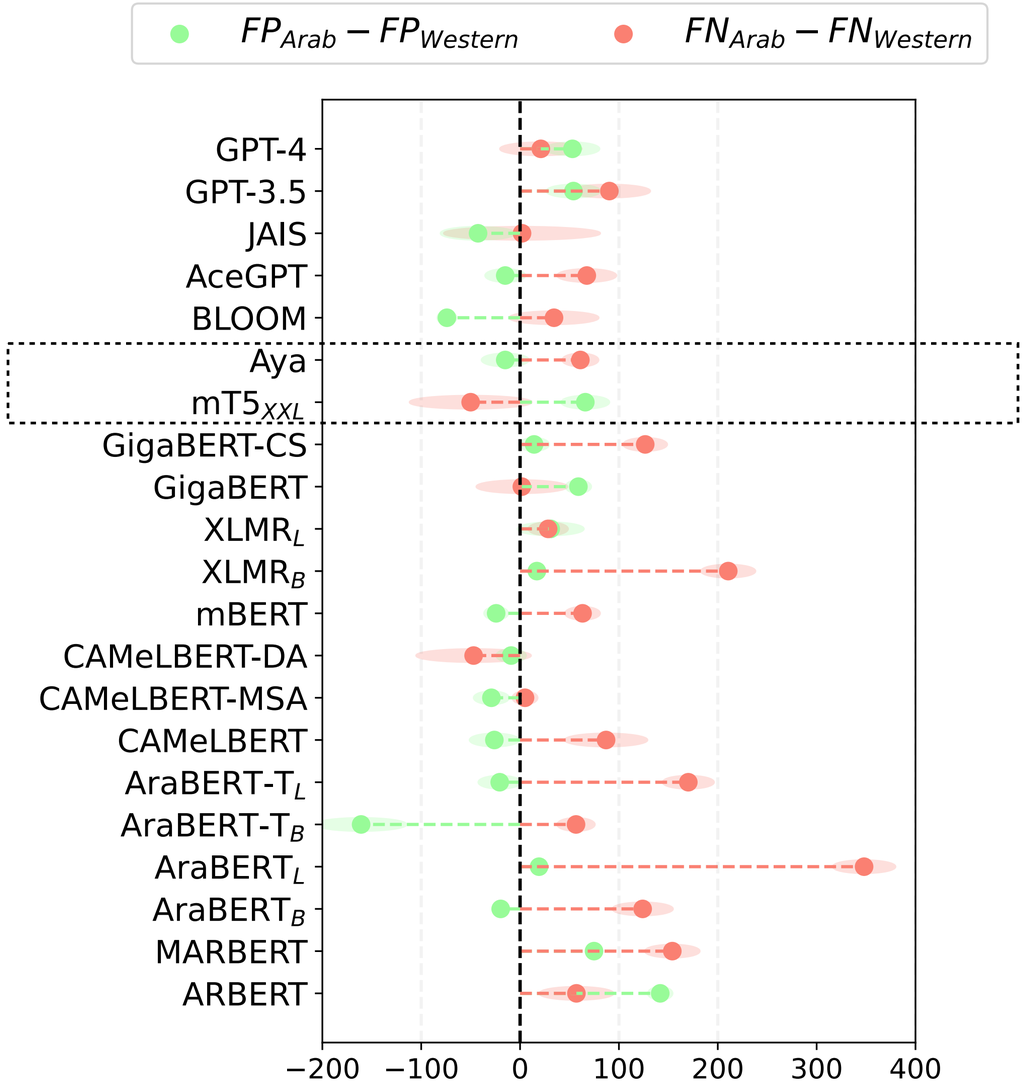
$$CBS = \frac{1}{\mathrm{N\,M\,K}} \sum_{i,j,k} \mathbb{I}[P_{[MASK]}(b_j \mid t_k) > P_{[MASK]}(a_i \mid t_k)]$$

# CAMeL — What about Sentiment?



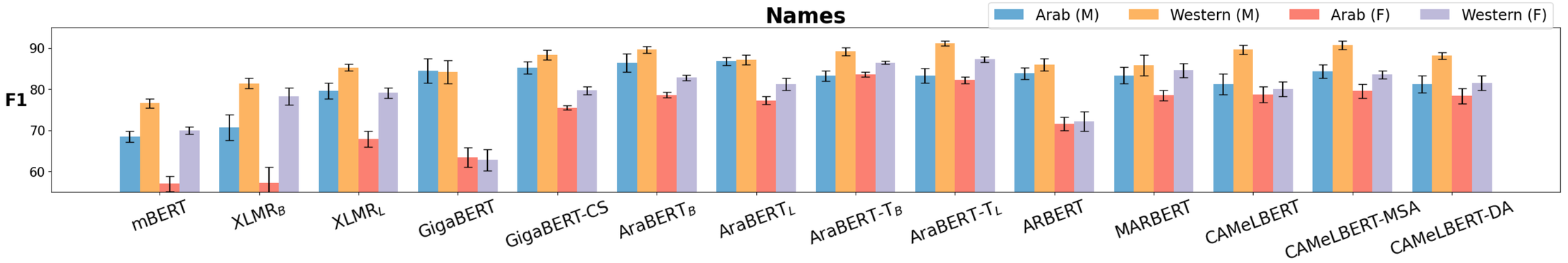

**Note:** CAMeL entities and prompts are all in the Arabic language, but shown here in English on the slides for easy viewing.

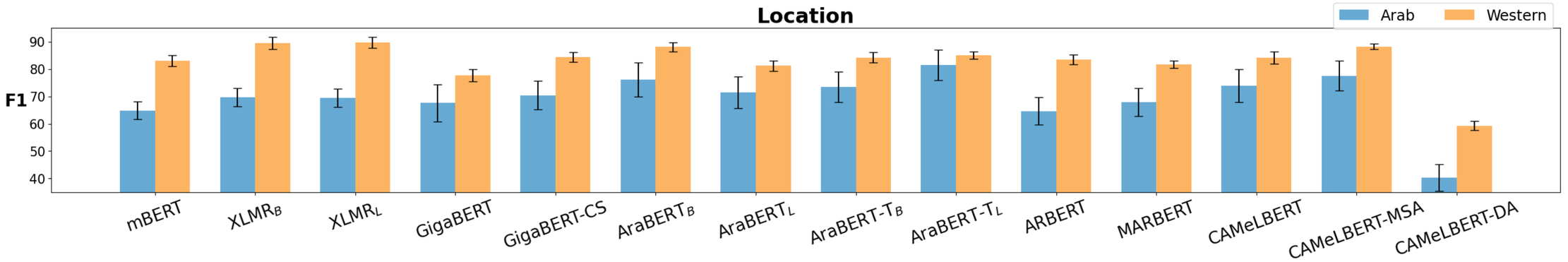
# CAMeL — more false negatives for Arabic entities



# CAMeL — What about NER of different entities?

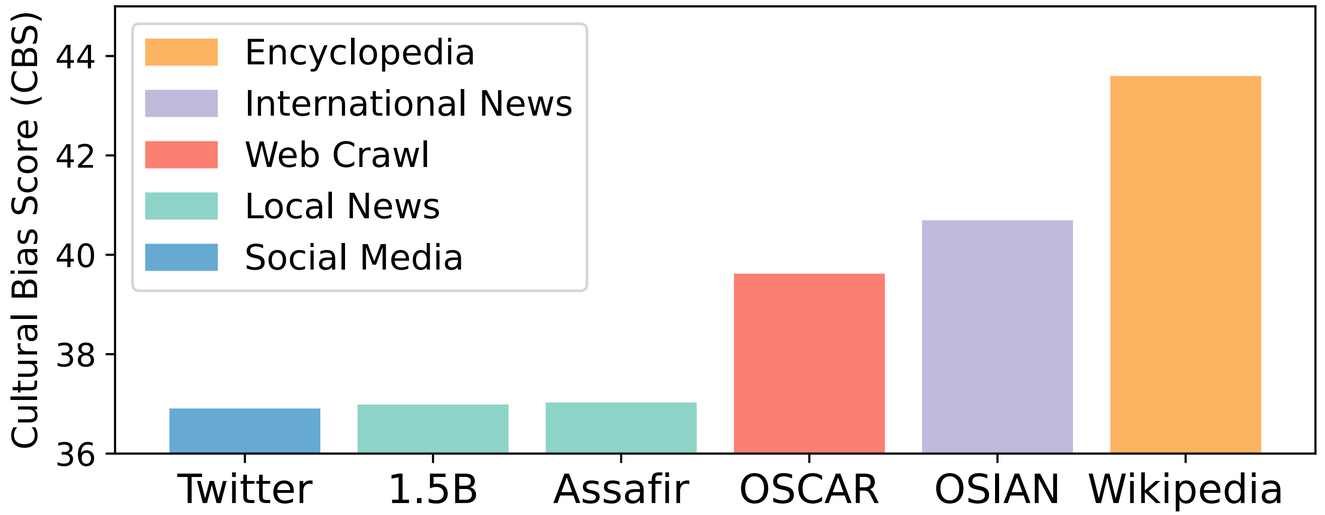
NER taggers are better at recognizing the Western person/location names than the Arab ones.



# CAMeL — What 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].”*

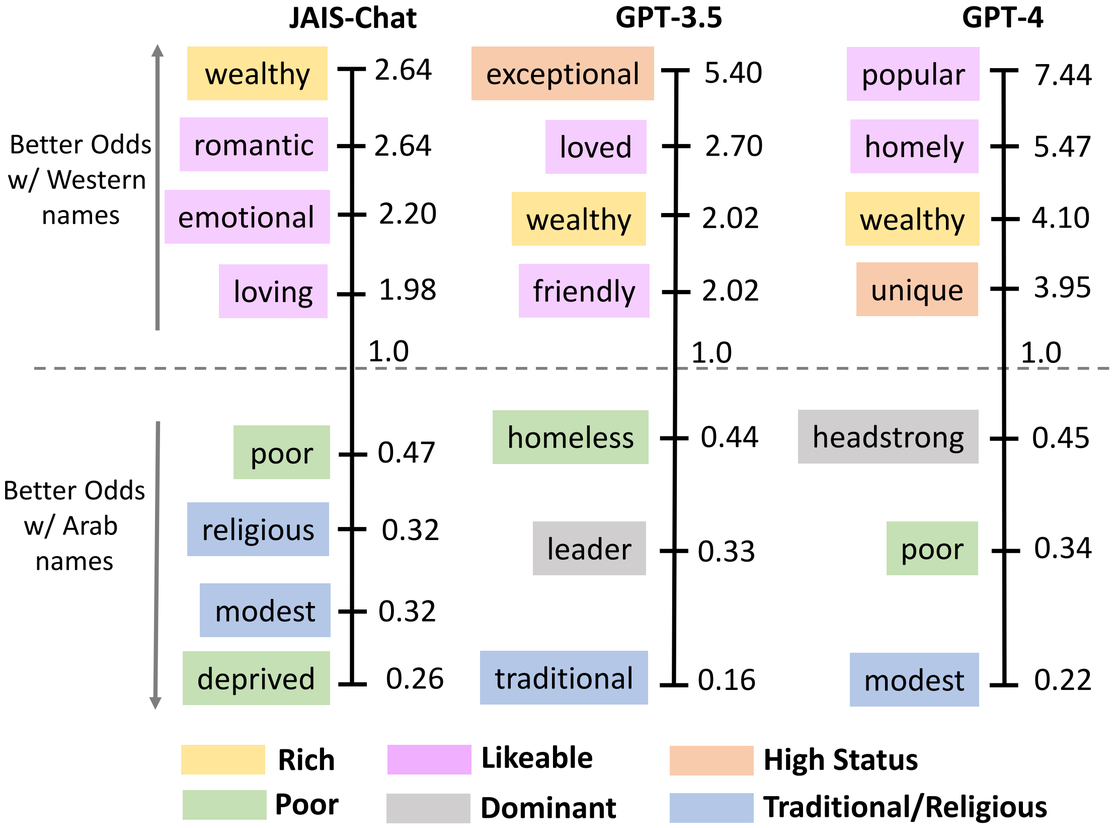
| **GPT-4** |
| --- |
| نشأ العاص في أسرة فقيرة و متواضعة وكانت الحياة بالنسبة له معركة يومية من أجل البقاء |
| (Al-Aas grew up in a poor and modest family where life was a daily battle for survival) |
| كان إيمرسون مشهورا بين أهل البلدة لذكائه الحاد ونظرته الثاقبة للأمور |
| (Emerson was popular in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

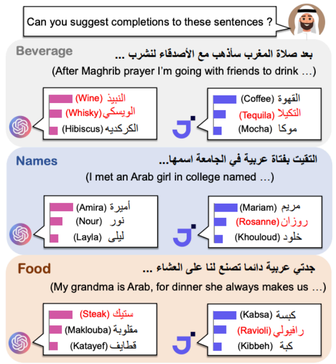
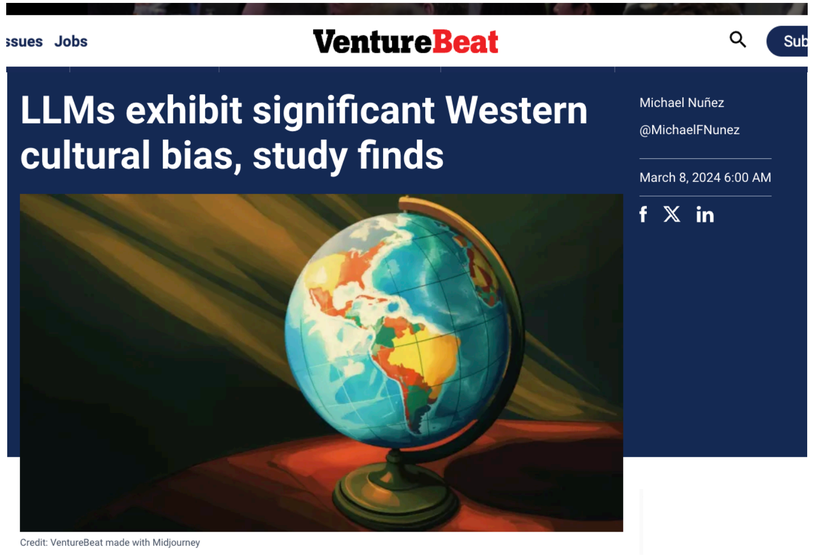



Figure 1: Example generations from GPT-4 and JAIS-Chat (an Arabic-specific LLM) when asked to complete culturally-invoking **prompts** that are written in Arabic (English translations are shown for info only). LMs often generate entities that fit in a Western culture (red) instead of the relevant Arab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

## Press Coverage




Credit: VentureBeat made with Midjourney

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

# Today’s talk —

## 1. Constrained Decoding

CODEC


(Le et al., ICLR 2024)

Decoding algorithms for better performance on non-English languages.

## 2. Cultural Bias

**CAMEL**

**(Naous et al., ACL 2024)**

Measuring implicit cultural bias in LLMs and pre-training data.

## 3. Neologism

NeoBench



(Zheng et al., ACL 2024)

How well LLMs can handle newly emerging words?

# Today’s talk —

## 1. Constrained Decoding

CODEC


(Le et al., ICLR 2024)

Decoding algorithms for better performance on non-English languages.

## 2. Cultural Bias

CAMEL



(Naous et al., ACL 2024)

Measuring implicit cultural bias in LLMs and pre-training data.

## 3. Neologism

**NeoBench**

**(Zheng et al., ACL 2024)**

How well LLMs can handle newly emerging words?

# Evaluating Robustness of Large Language Models with Neologisms (NeoBench)


Jonathan Zheng

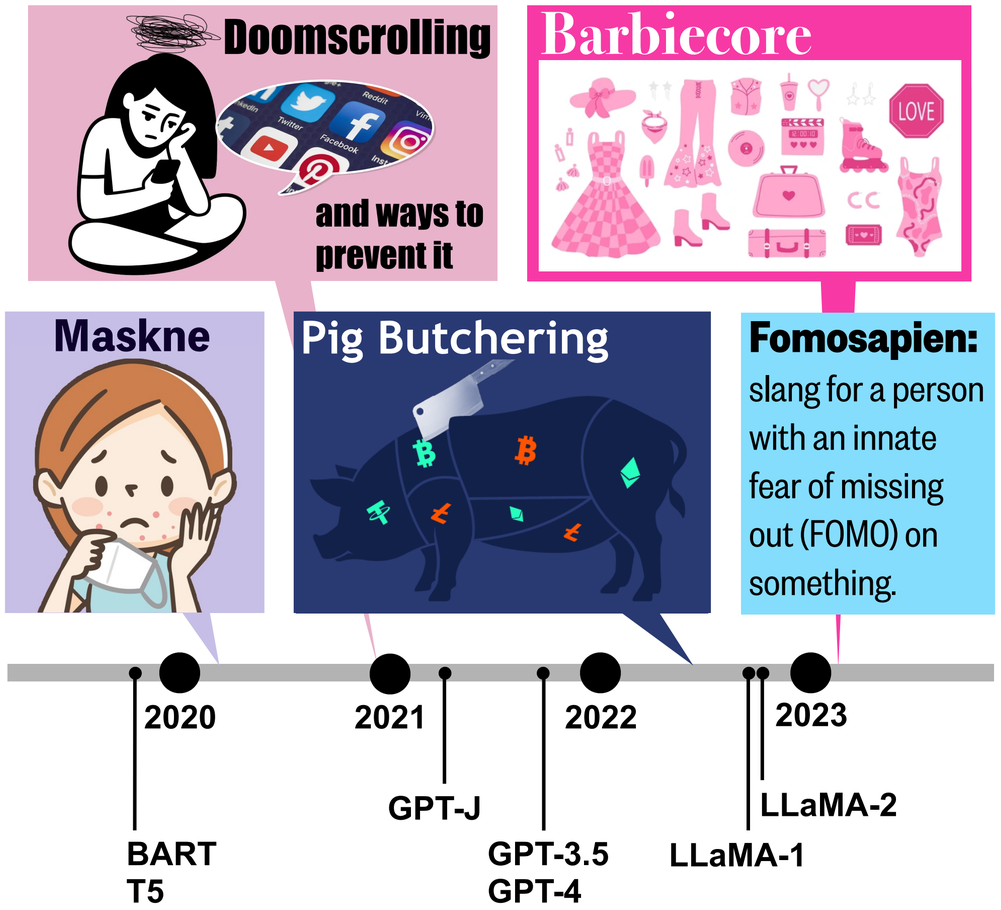
Alan Ritter

Wei Xu


A better technical solution for marker-based label projection

# NeoBench — evolving human languages

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




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

# NeoBench — human evaluation on translation

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

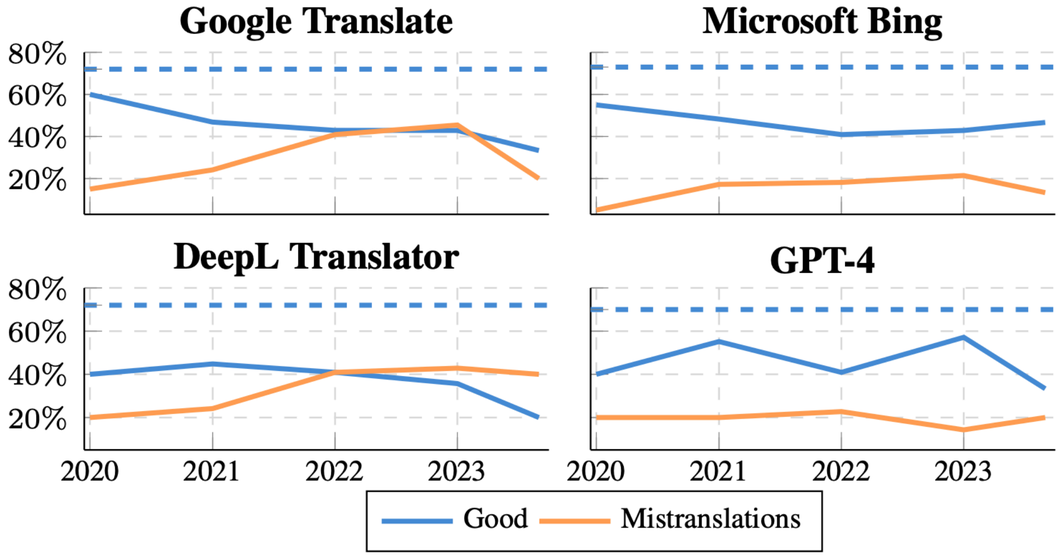



Figure 3: Percentage of good translations and mistranslations of neologism sentences over time. The dashed line represents the percentage of good translations achieved on non-neologism sentences.

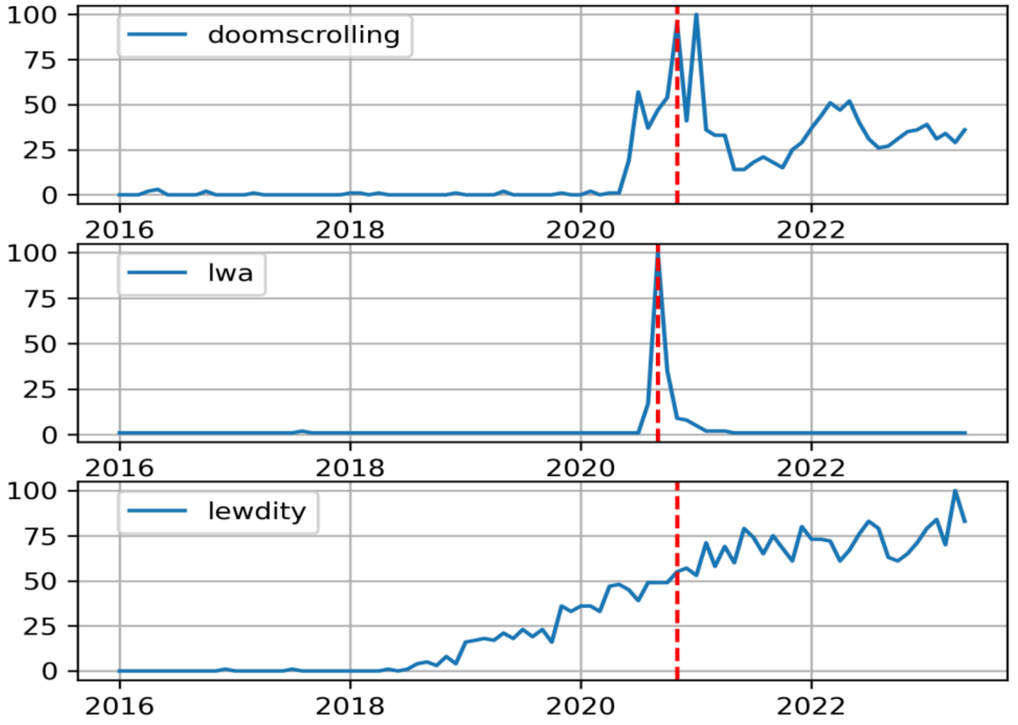
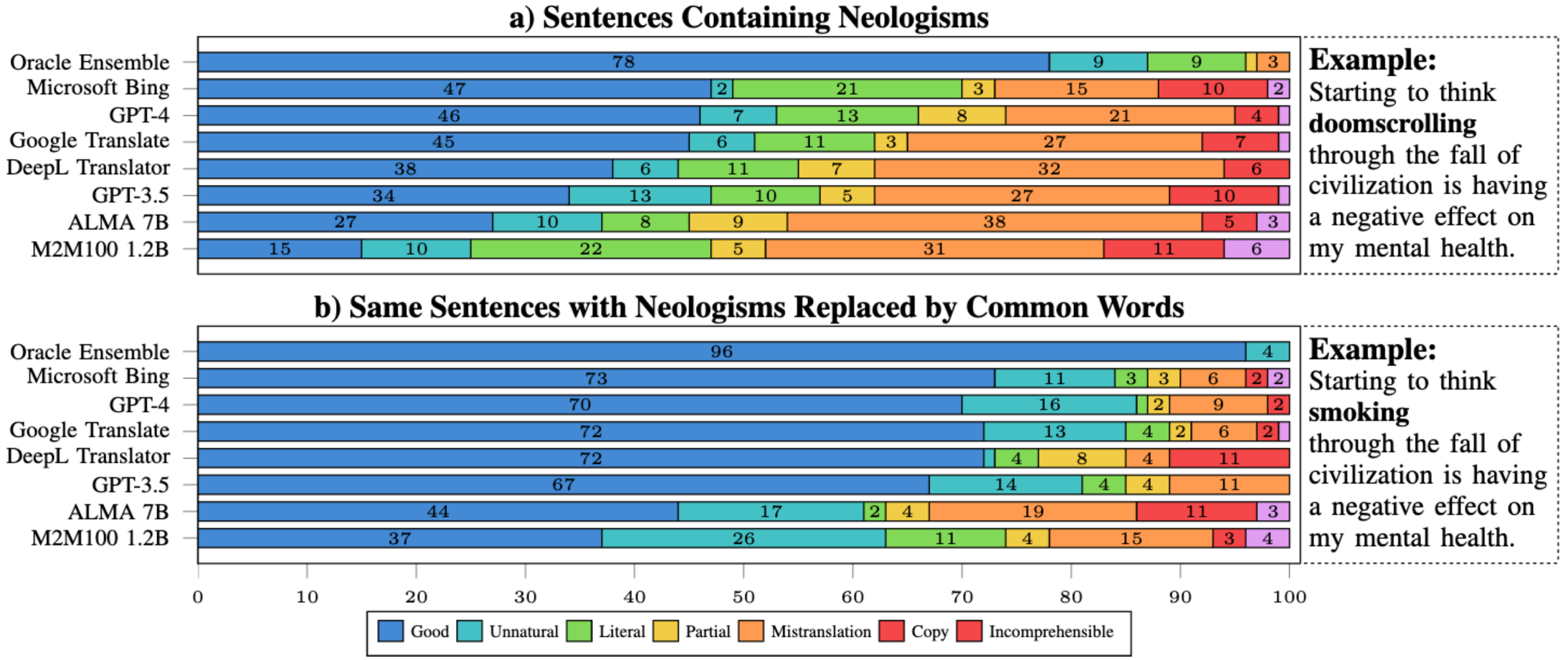



Figure 4: Example Google Trend lines measuring neologism prevalence. The dashed line estimates the date a neologism becomes popular while not yet conventional.

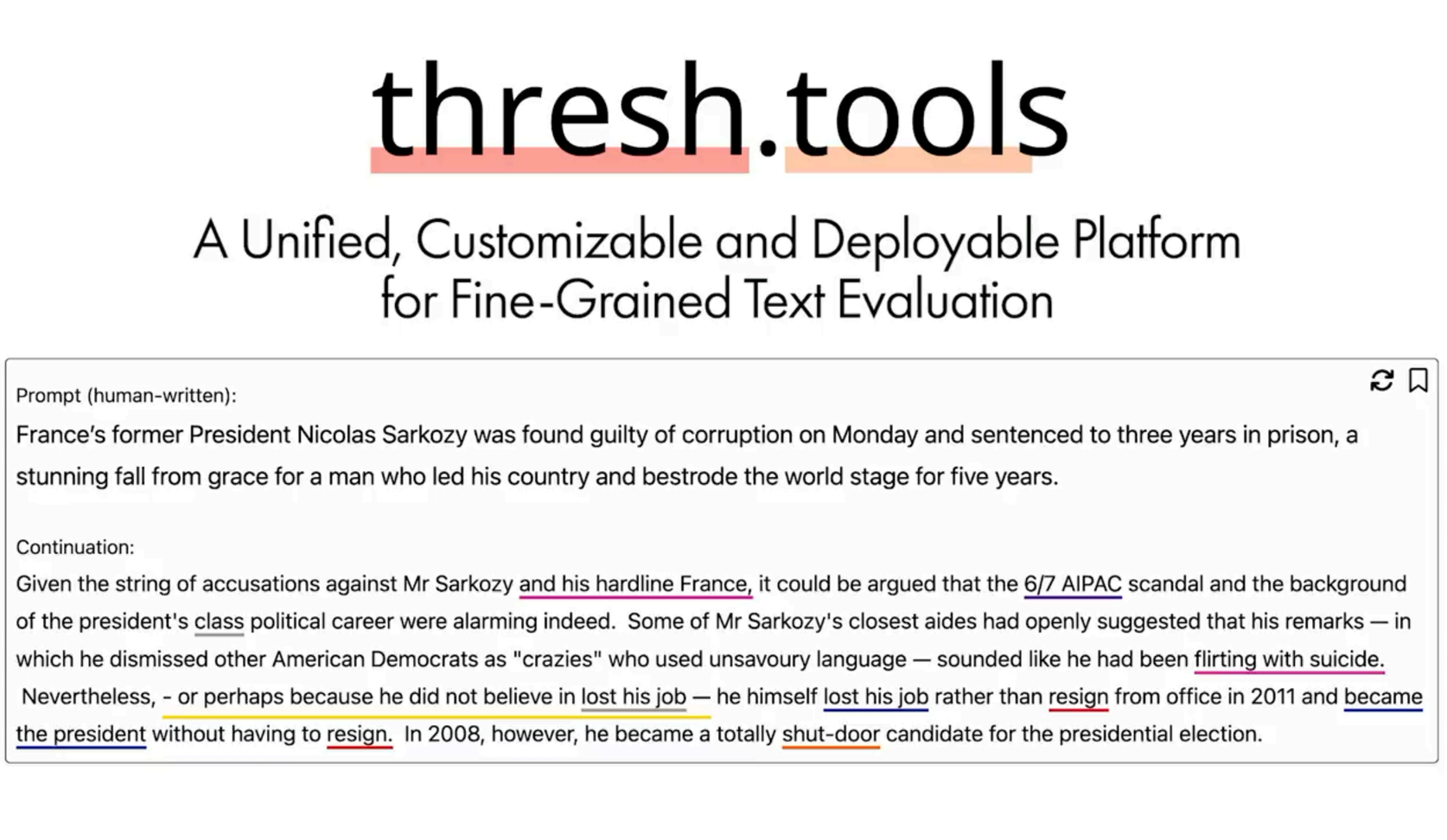
# NeoBench — human evaluation on translation

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



# Thresh — Fine-grained Text Evaluation Tool

https://github.com/davidheineman/thresh




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

# NeoBench — perplexity, Cloze QA, definition

| | |
|---|---|
| **a) Definition Generation Output Examples** | |
| **Stablecoin** | **Reference Definition:** *A stablecoin is a type of cryptocurrency where the value of the digital asset is supposed to be pegged to a reference asset, which is either fiat money, exchange-traded commodities, or another cryptocurrency.* |
| | **Model Output (Correct):** *Stablecoins are cryptocurrencies designed to maintain a stable value, typically by pegging their value to a specific asset or basket of assets, such as the US dollar, gold, or a combination of assets.* |
| **Angel Shot** | **Reference Definition:** *An angel shot is a code to inform a bartender that a customer is not safe and needs assistance.* |
| | **Model Output (Incorrect):** *An angel shot is a cocktail made with whiskey and cream, served in a shot glass.* |
| **b) Machine Translation Output Examples** | |
| **Longcovid** | **Input:** Each reinfection increases the risk of **longcovid**, hospitalization, & death. |
| | **Model Output (Correct):** 每次再感染都会增加长新冠病毒、住院和死亡的风险。<br>*(Every reinfection increases the risk of **long COVID**, hospitalization, and death.)* |
| | **Human Translation:** 每一次新冠感染都会提高出现后遗症、住院治疗，甚至死亡的风险。<br>*(Each COVID-19 infection increases the risk of developing **sequelae**, hospitalization, and even death.)* |
| **Doomscrolling** | **Input:** Starting to think **doomscrolling** through the fall of civilization is having a negative effect on my mental health. |
| | **Model Output (Incorrect):** 开始认为在文明的衰落中滚动的厄运对我的心理健康产生了负面影响。<br>*(Start to think that the **doom rolling** in the decline of civilization is having a negative impact on my mental health.)* |
| | **Human Translation:** 开始觉得, 刷关于文明衰败的负能量新闻对我的心理健康产生了负面影响。<br>*(Starting to feel that **scrolling** through negative news about the decline of civilization is having a negative impact on my mental health.)* |

Table 3: Example model definitions and translations for NEO-BENCH tasks. *"Doomscrolling"* is the act of spending an excessive amount of time reading negative news online. (English translations are shown for information only.)

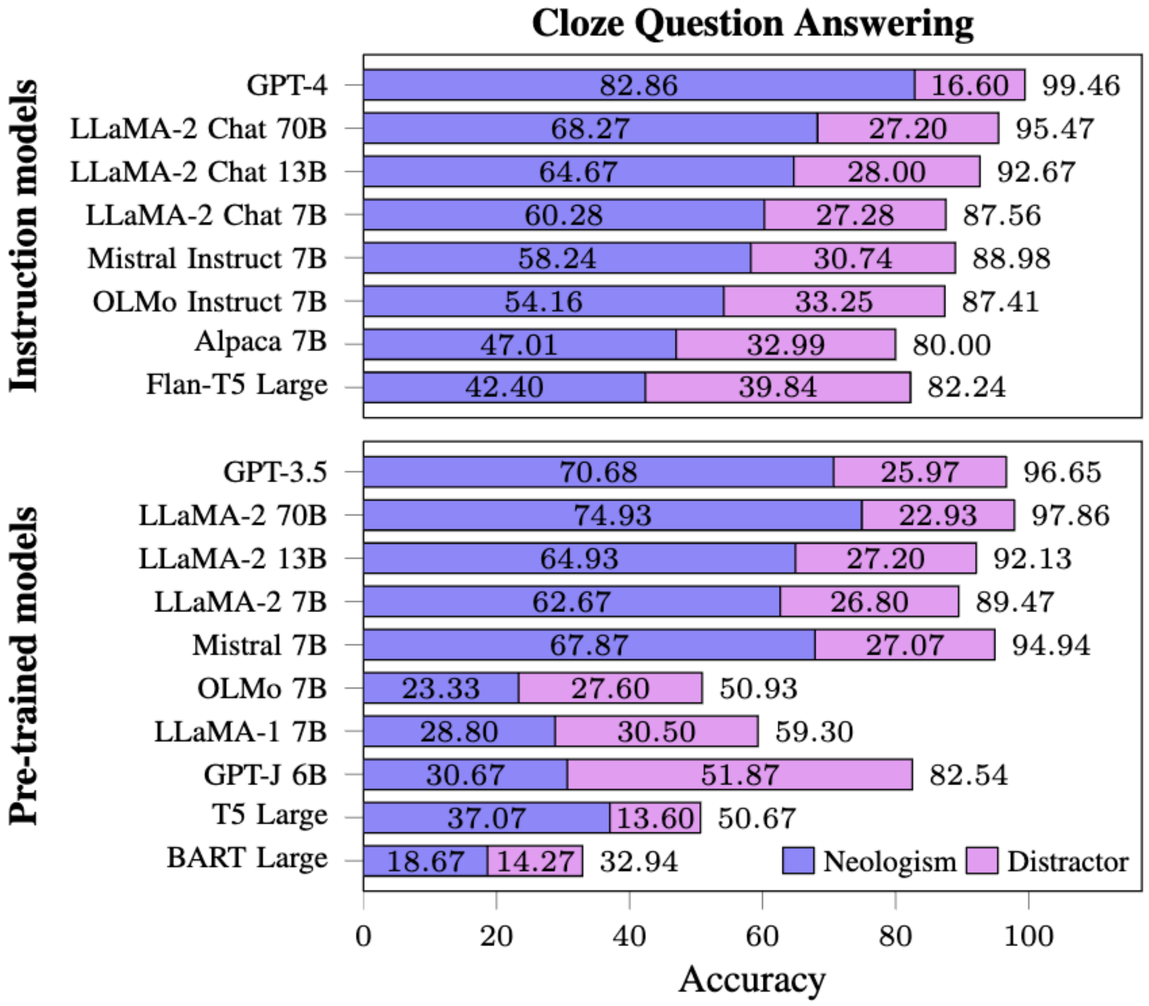
# NeoBench — automatic eval on translation

Automatic evaluation metrics do not show good system-level correlations with human evaluation.

| **Model** (human rank) | | **Reference-Based Metrics** | | | | **Reference-Free Metrics** | | |
|---|---|---|---|---|---|---|---|---|
| | | **BLEU**↑ | **COMET**↑ | **MX-23**$_{XXL}$↓ | **MX-23**$_{XL}$↓ | **COMET**$_{KIWI}$↑ | **MX-QE**$_{XXL}$↓ | **MX-QE**$_{XL}$↓ |
| Bing Translator | (1) | 0.452 (2) | 0.825 (5) | 2.419 (6) | 2.343 (6) | 0.788 (5) | 1.679 (5) | 2.246 (5) |
| GPT-4 | (2) | 0.446 (3) | **0.854** (1) | **1.550** (1) | **1.793** (1) | 0.793 (3) | 1.432 (3) | 2.089 (3) |
| Google Translate | (3) | **0.507 (1)** | 0.853 (2) | 1.825 (4) | 1.945 (4) | 0.800 (2) | 1.429 (2) | **1.940** (1) |
| DeepL Translator | (4) | 0.406 (4) | 0.842 (3) | 1.775 (3) | 1.901 (3) | **0.807** (1) | **1.260** (1) | 1.944 (2) |
| GPT-3.5 | (5) | 0.399 (5) | 0.841 (4) | 1.705 (2) | 1.796 (2) | 0.792 (4) | 1.467 (4) | 2.157 (4) |
| ALMA 7B (LLaMA-2) | (6) | 0.285 (7) | 0.801 (6) | 2.382 (5) | 2.251 (5) | 0.746 (6) | 2.038 (6) | 2.462 (6) |
| M2M100 1.2B | (7) | 0.337 (6) | 0.776 (7) | 3.454 (7) | 3.142 (7) | 0.745 (7) | 2.821 (7) | 2.976 (7) |
| **Spearman's** $\rho$ | | 0.244 | 0.445 | 0.457 | 0.380 | **0.491** | 0.451 | 0.445 |

# NeoBench — perplexity, Cloze QA, definition

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

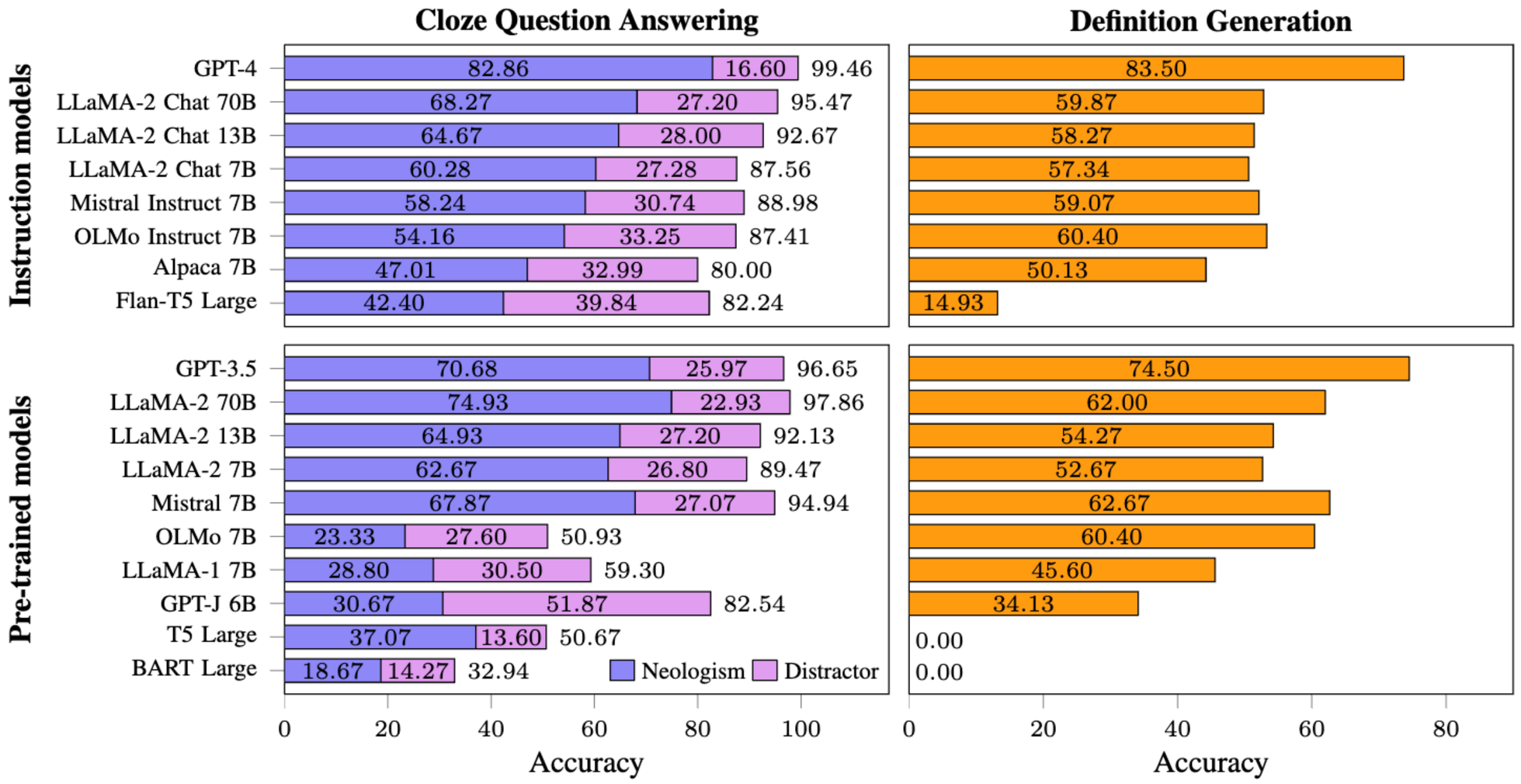



| **Neologism: *doomscrolling*** | |
|---|---|
| *The silver lining of this website no longer functioning as an even vaguely reliable information source is that ___ has basically been completely undermined. It wouldn't even work now since everything is too geared to outrage clickbait and actual reporting has disappeared, so there is no point staying on the app.* | |
| *a) misinformation* | *b) surfing* |
| ***c) doomscrolling*** | *d) lying* |
| *e) gaming* | ***f) anti-productivity (distractor)*** |

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

# NeoBench — perplexity, Cloze QA, definition

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



# Today’s talk — let’s wrap up!

## 1. Constrained Decoding

**CODEC**


**(Le et al., ICLR 2024)**

Decoding algorithms for better performance, data argumentation, non-English languages.

## 2. Cultural Bias

**CAMEL**

**(Naous et al., ACL 2024)**

Systematically analyze implicit cultural bias in LLMs and pre-training data.

## 3. Temporal Shift

**NeoBench**

**(Zheng et al., ACL 2024)**

How well LLMs can handle newly emerging (“unseen”) words?

# Conclusions

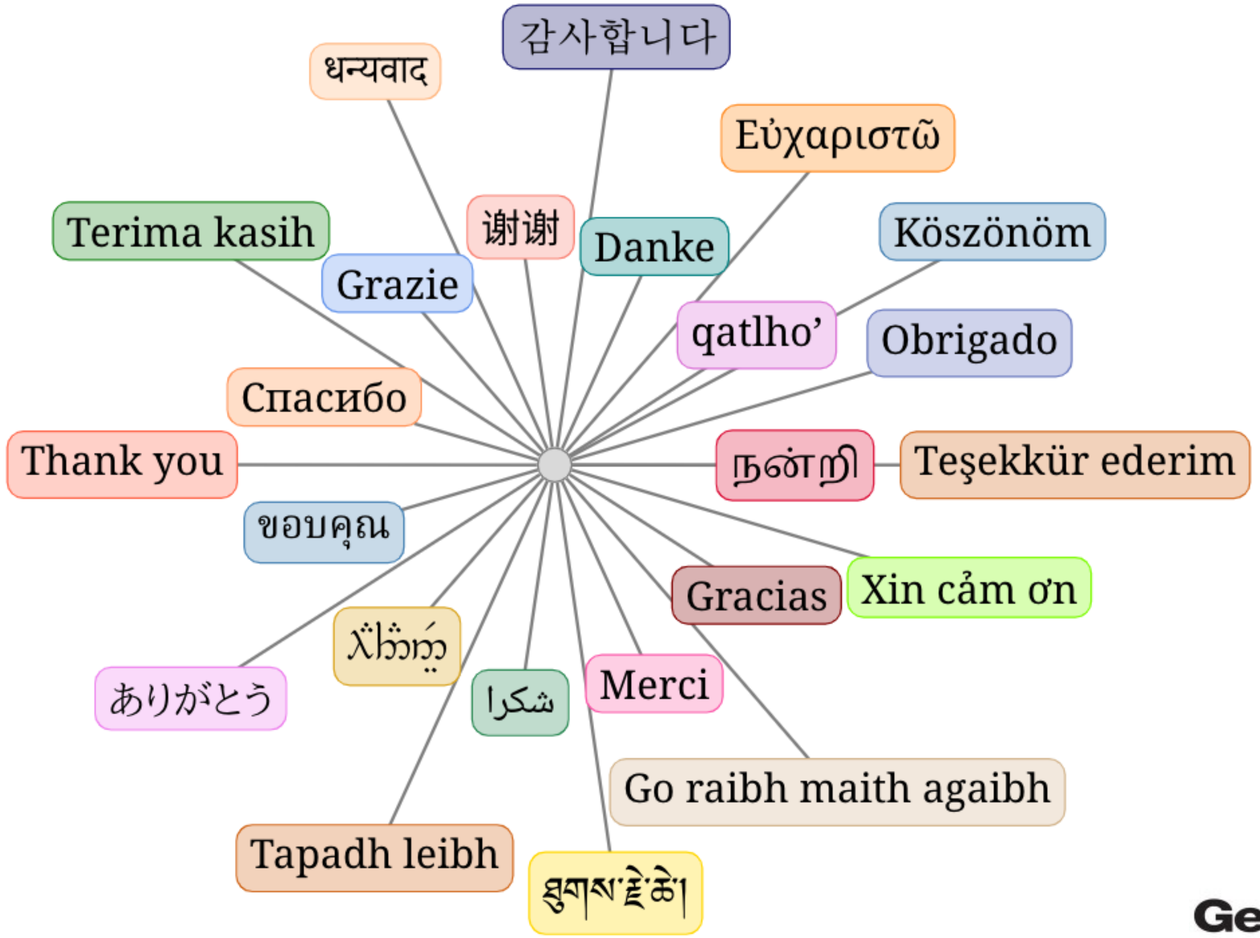
1. We need not only multilingual LLMs, but also multicultural, multi-domain LLMs.

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

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

# Thank you!

https://cocoxu.github.io/



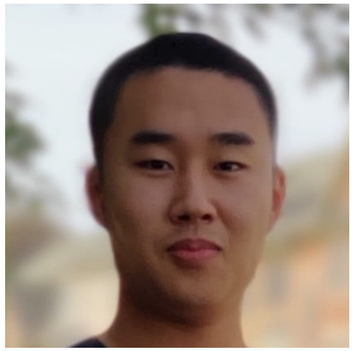

(image credit: Overleafl)


(image credit: Georgia Tech)

Fine-grained quality evaluation of LLM-generated text

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


David Heineman



Yao Dou

Wei Xu

Georgia Tech

# 🌾Thresh — good or bad LLM generations

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

# 🌾Thresh — good or bad LLM generations

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

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

# 🌾Thresh — good or bad LLM generations

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

| | |
|---|---|
| **Original** | It was originally thought that the debris thrown up by the collision filled in the smaller craters. |
| **(Sulem et al., 2018)** | It was originally thought that the debris thrown up by the. Collision filled in the smaller craters |
| **(Maddela et al., 2020)** | ~~It was originally thought that~~ the debris thrown up by the collision filled in the smaller craters. |
| **GPT-3.5, 2022** | It was believed that the smaller craters were filled in by debris from the collision. |
| **Human** | The smaller craters were originally thought to be filled by collision debris. |

# 🌾Thresh — good or bad LLM generations

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

**Paraphrase** **Deletion** **Insertion** || **Split**

Complex Sentence:

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

Simplification by GPT-4:

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

Can you spot the errors that GPT-4 made?

# 🌾Thresh — good or bad LLM generations

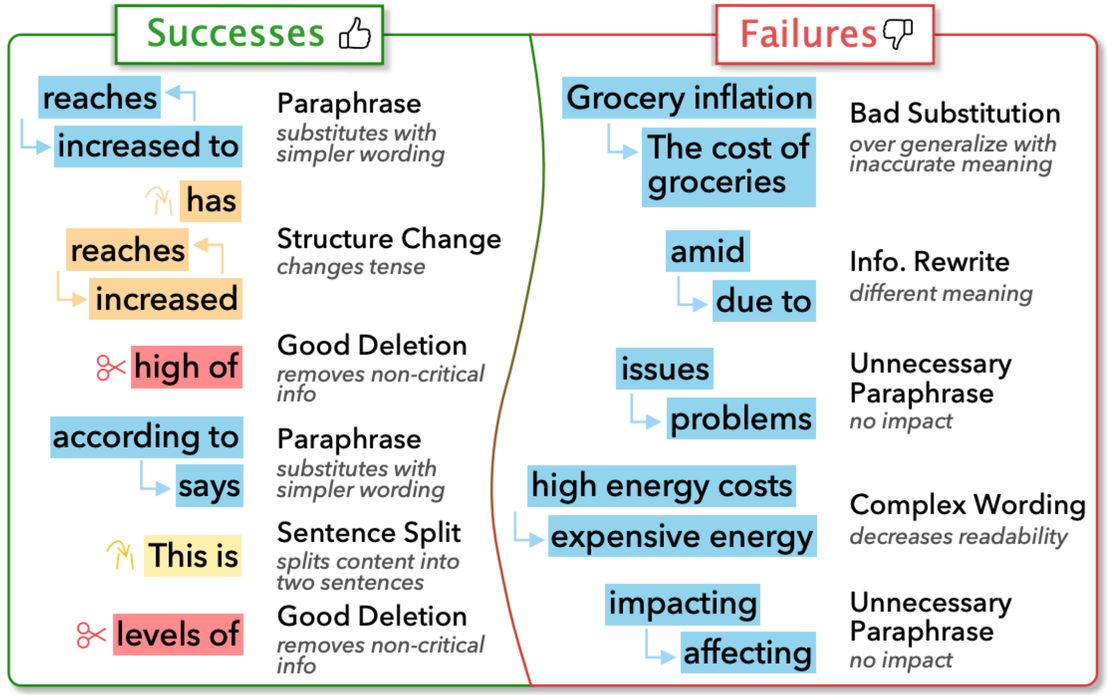
Here is another example of text simplification. GPT-4 rewrites complex text into simpler language.

**Complex Sentence:**

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

**Simplification by GPT-4:**

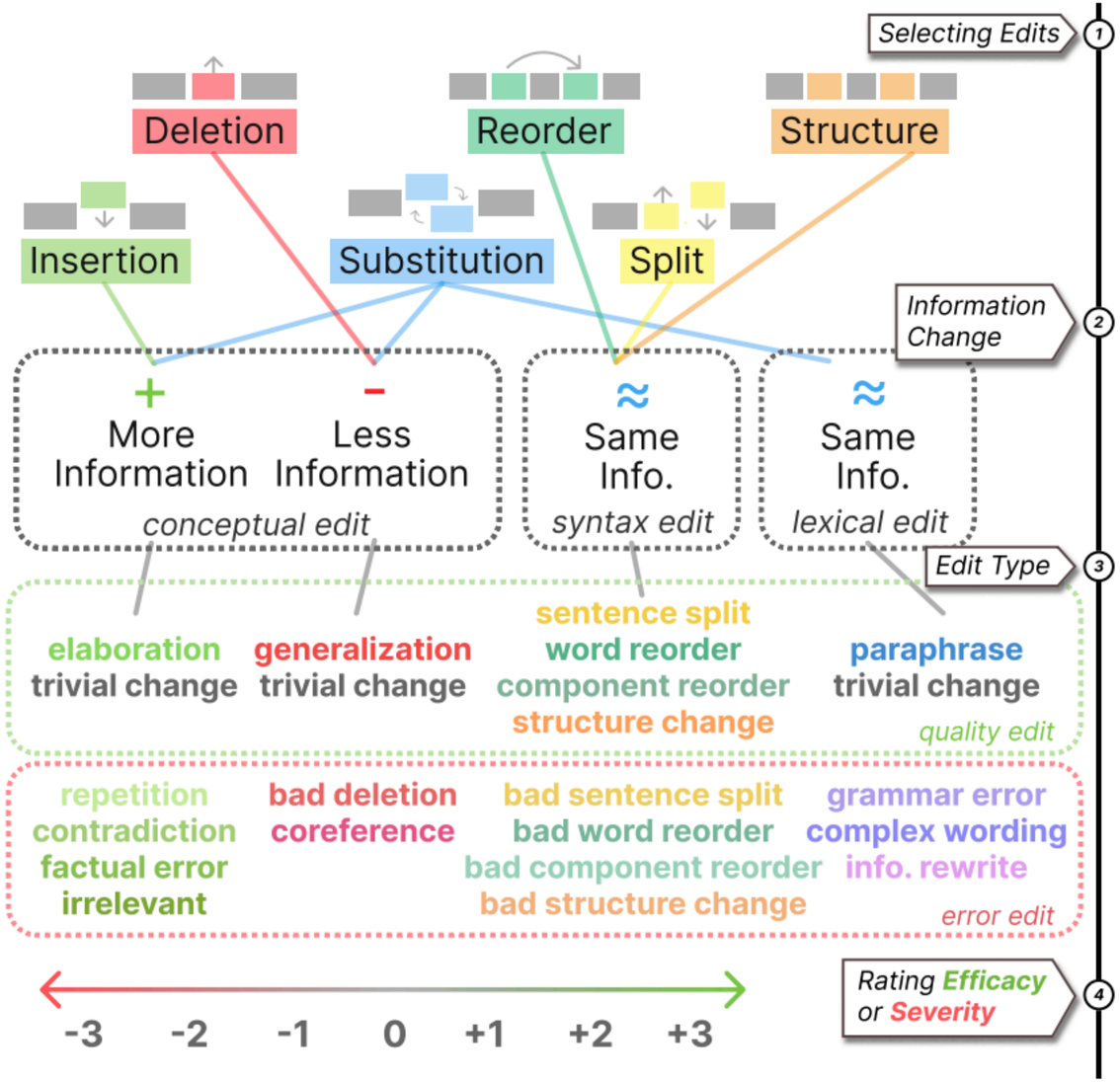
The cost of groceries in the United Kingdom has increased to a record 17.1%, says market research group Kantar Worldpanel. || This is due to high inflation, supply chain problems, and expensive energy affecting the economy.




Errors in LLM-generated texts can be difficult to capture

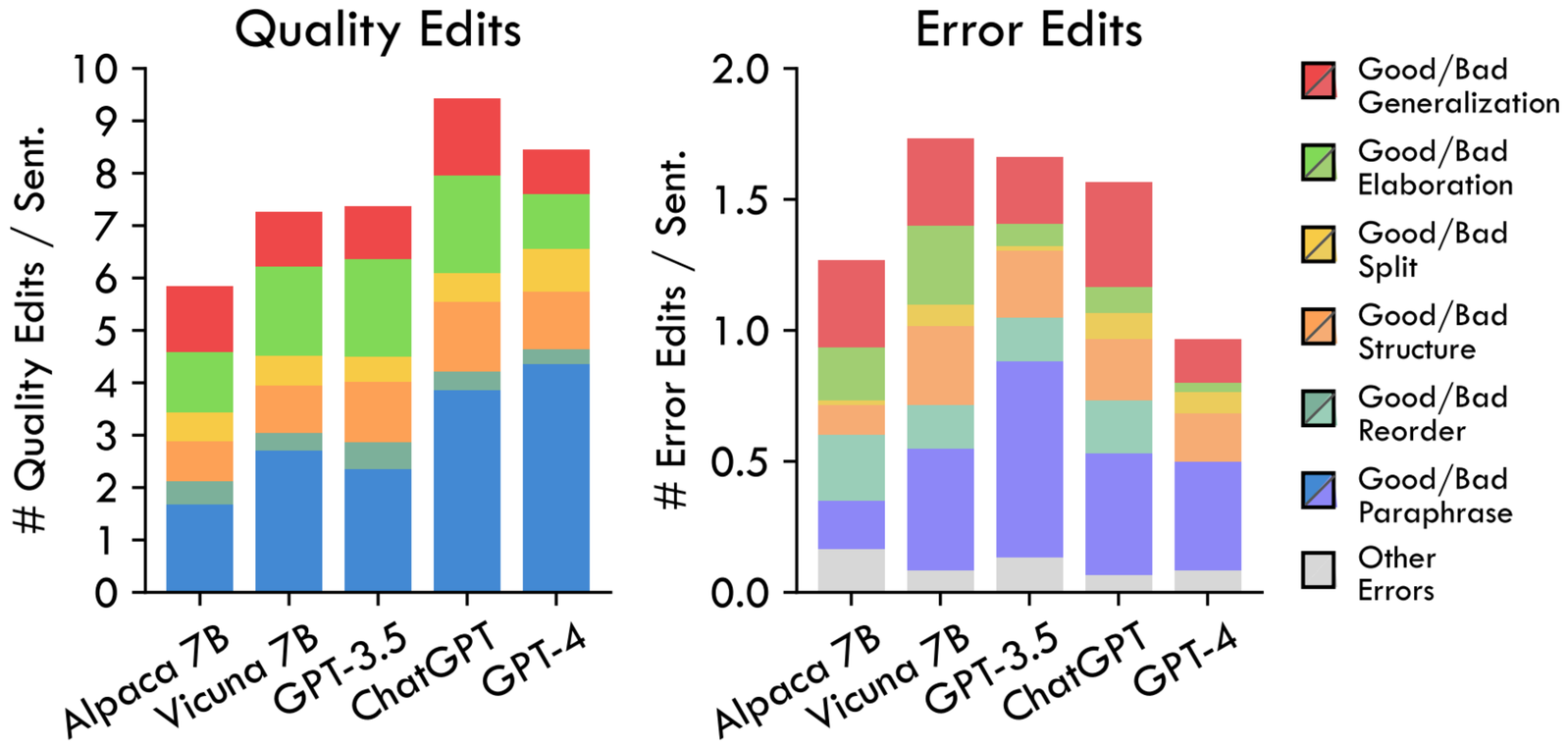
# 🌾Thresh — typology for edit-level evaluation

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



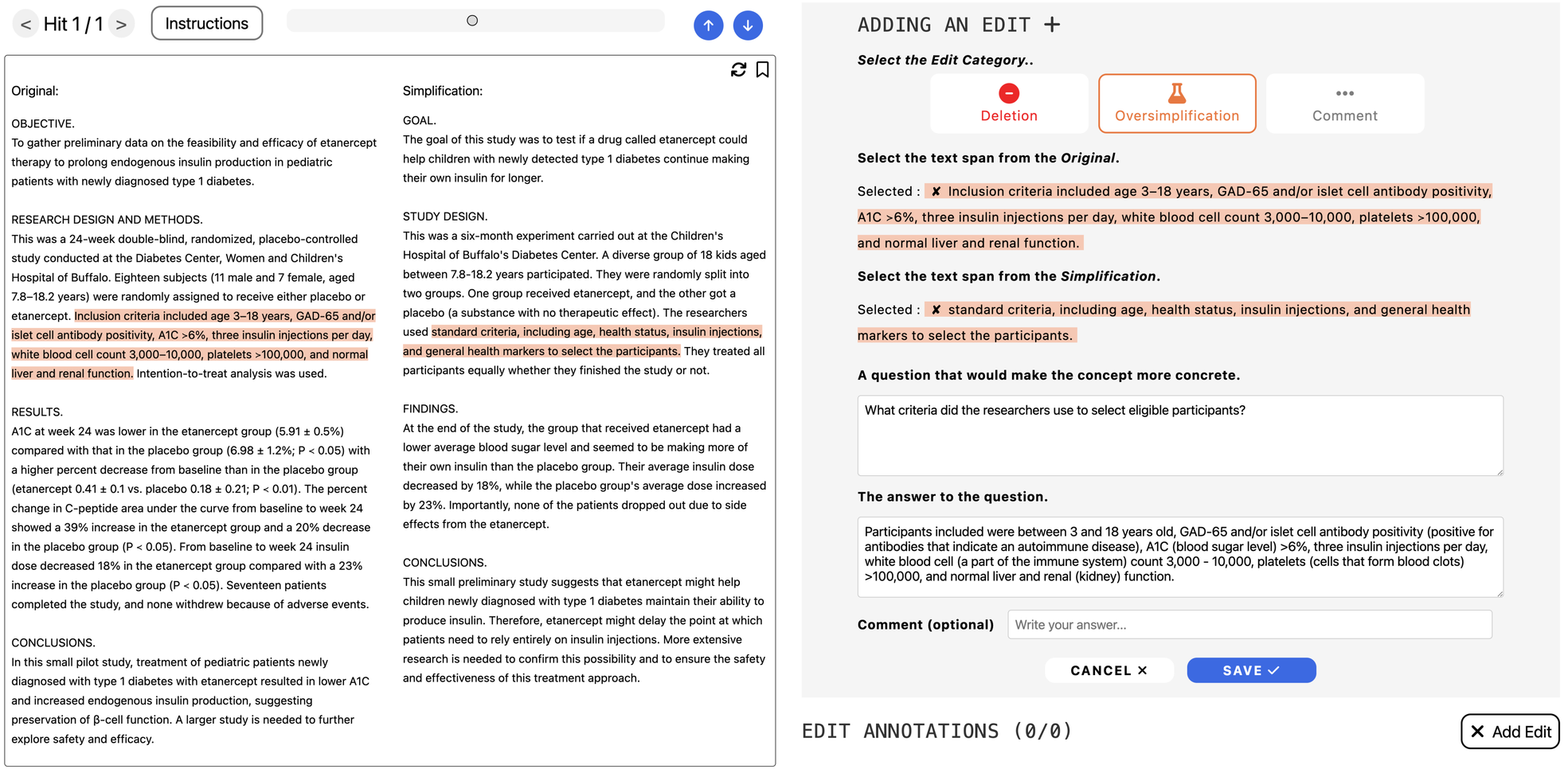
# Thresh — analysis of LLM-generated text

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



# 🌾Thresh — support document-level evaluation

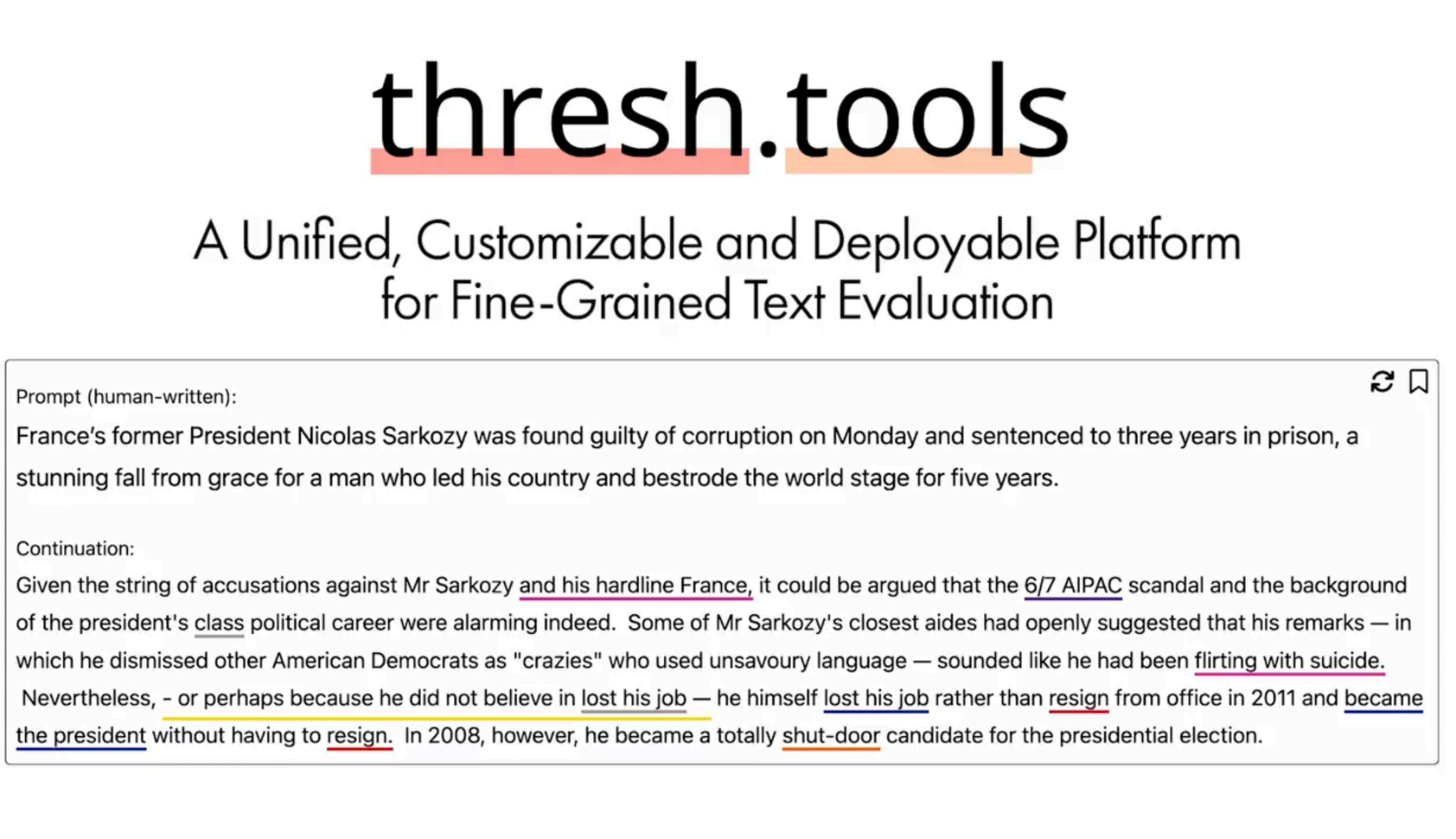
Here is an example where Thresh is extended to evaluate document-level medical text simplification in the form of question-answering pairs.




Jan Trienes, Sebastian Joseph, Jörg Schlötterer, Christin Seifert, Kyle Lo, Wei Xu, Byron C. Wallace, Junyi Jessy Li.
“InfoLossQA: Characterizing and Recovering Information Loss in Text Simplification” (ACL 2024)

# Thresh — Fine-grained Text Evaluation Tool

https://github.com/davidheineman/thresh




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

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

# Conclusions

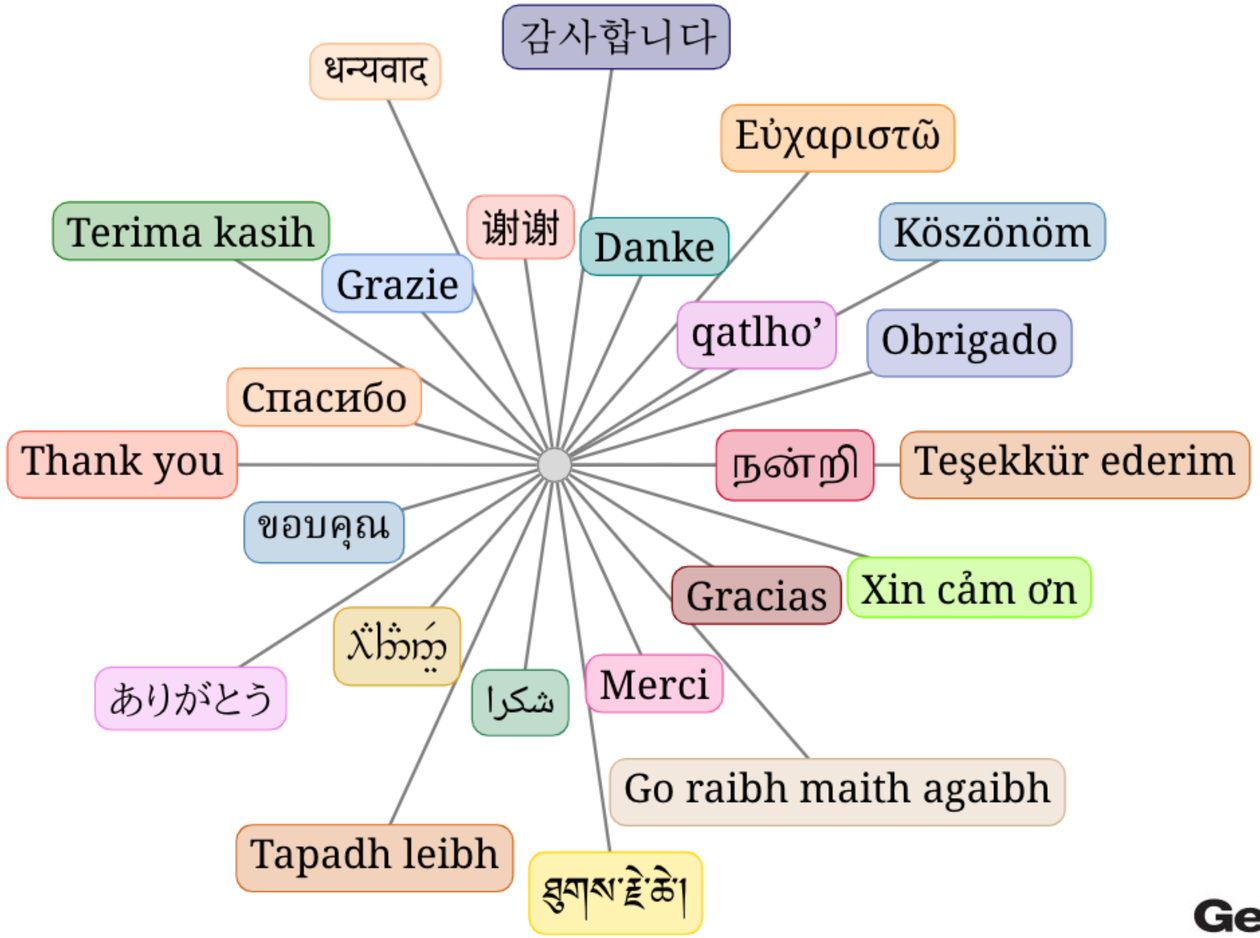
1. We need not only multilingual LLMs, but also multicultural, multi-domain LLMs.

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

3. 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: Overleafl)


(image credit: Georgia Tech)

# Backup slides for EasyProject

# Experiments

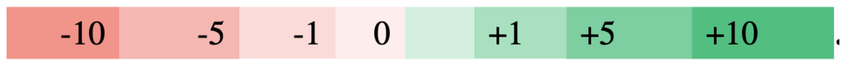
Three Datasets:

- 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

Δ from fine-tune XLM-R on English baseline




| | *en* → Lang. | Fine-tune$_{en}$ | | NLLB+Word Align | | | NLLB+Markers | | GMT+Word Align | | | GMT+Markers | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Ref. | XLM$_R$ | QAali. | Awes. | Awes$_{ft}$ | XML | EProj. ($\Delta_{XLM_R}$) | QAali. | Awes. | Awes$_{ft}$ | 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) |
| | *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:

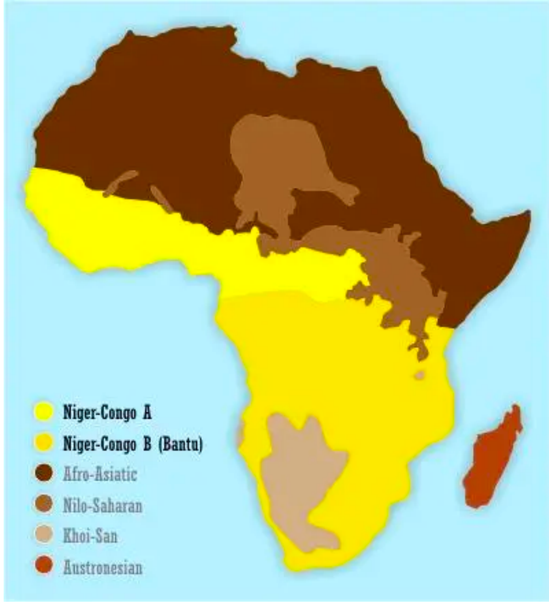
  mBERT, mDebertaV3, AfroXLMR, Glot500 — specialized for African languages

- Compare to a modified version of beam search with the constrained search space
- And more ….

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

# 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)

# Experiment Results

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

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

| Lang. | $FT_{en}$ | Awes-align | EasyProject | CODEC |
|---|---|---|---|---|
| Arabic | 44.8 | 48.3 | 45.4 | **48.4** |
| Chinese | 54.0 | 57.3 | **59.7** | 59.1 |
| AVG | 49.4 | 52.8 | 52.6 | **53.8** |

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

$FT_{En}$: Fine-tune on English data only

- Event Extraction: mT5-large
- Translation: NLLB

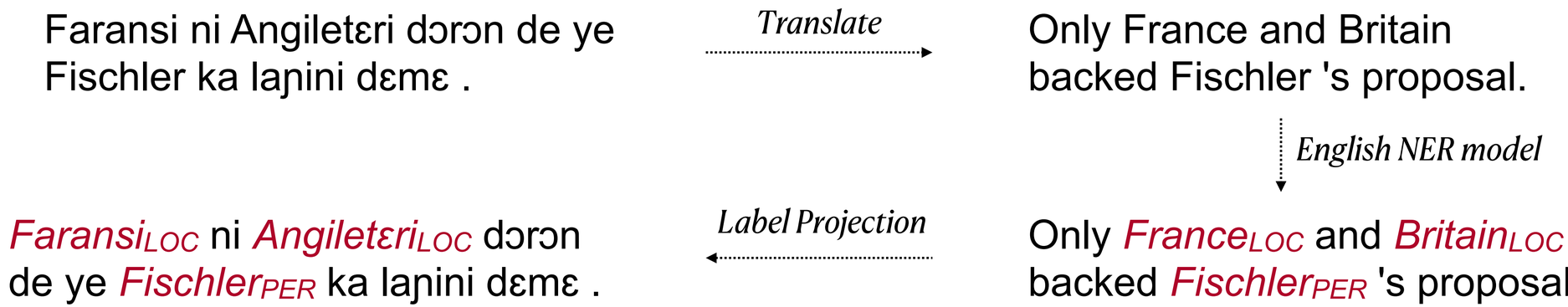
# Two Label Projection Setups

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)*



# 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**

| $FT_{En}$ | 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 | **70.4** (+16.0) |

*Awes-align*: Alignment-based approach | *EasyProject*: marker-based approach
*$FT_{En}$*: Fine-tune on English data only