

# DLAMA: A Framework for Curating Culturally Diverse Facts for Probing the Knowledge of Pretrained Language Models

Amr Keleg, Walid Magdy

University of Edinburgh

*a.keleg@sms.ed.ac.uk*

**Accepted to ACL 2023 (Findings)**



Institute for Language,  
Cognition and Computation



THE UNIVERSITY of EDINBURGH  
**informatics**



**Large Language Model (M)**  
can solve **(T)** better than other  
models.



**Large Language Model (M)**  
can solve **(T)** better than other  
models.

**Fair** evaluation needs

**BENCHMARKS**



# What?

# What?

Factual World Knowledge.

e.g.: (Edinburgh, **Country**, Scotland)

# What?

Factual World Knowledge.

e.g.: (Edinburgh, **Country**, **Scotland**)

# Why?

# What?

Factual World Knowledge.

e.g.: (Edinburgh, **Country**, **Scotland**)

# Why?

- Better **Reasoning** and Inference

# What?

Factual World Knowledge.

e.g.: (Edinburgh, **Country**, Scotland)

# Why?

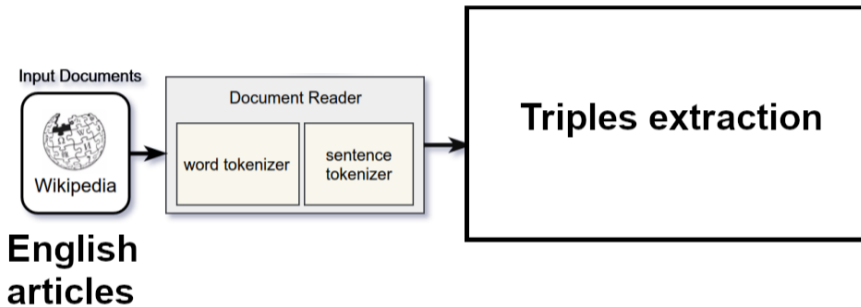
- Better **Reasoning** and Inference
- Replace **manual** curation of triples

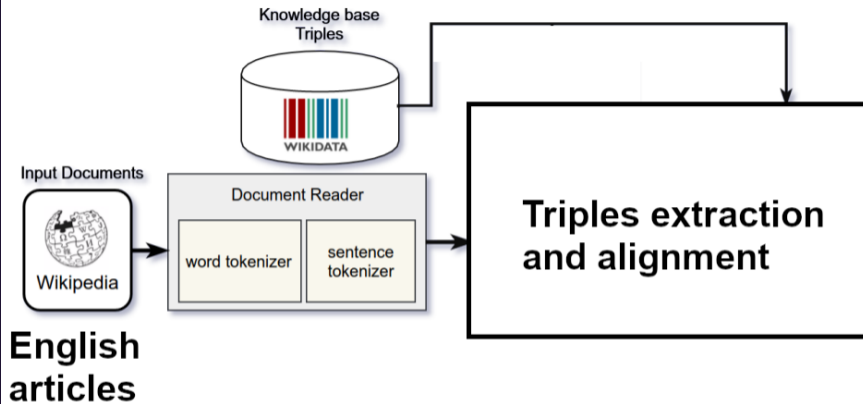


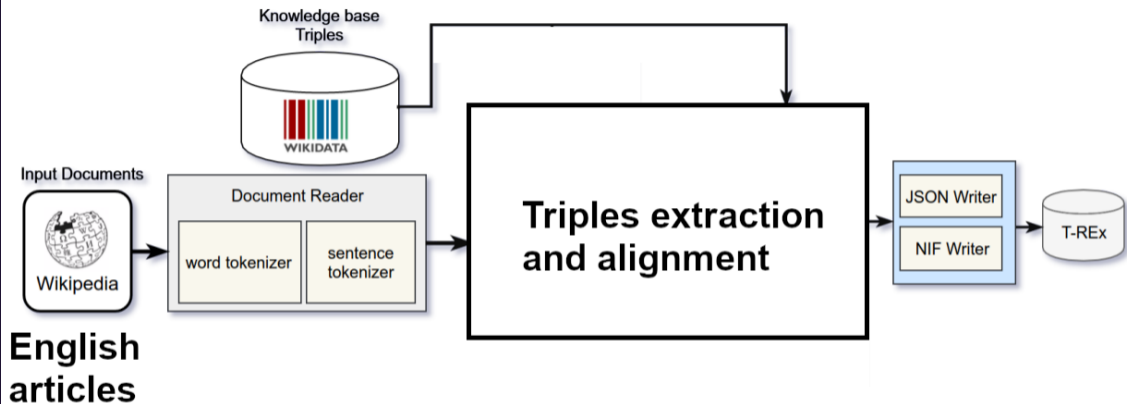
Current benchmarks are built for English then translated to other languages.

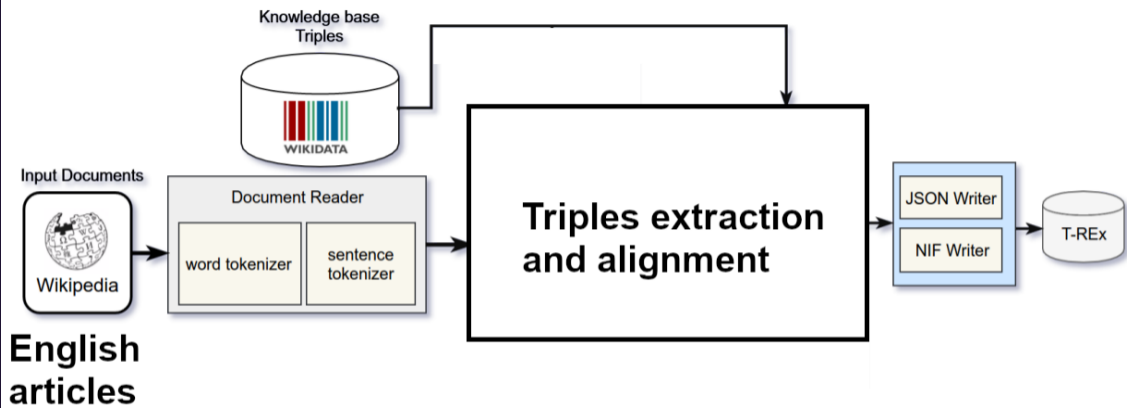
Knowledge base  
Triples



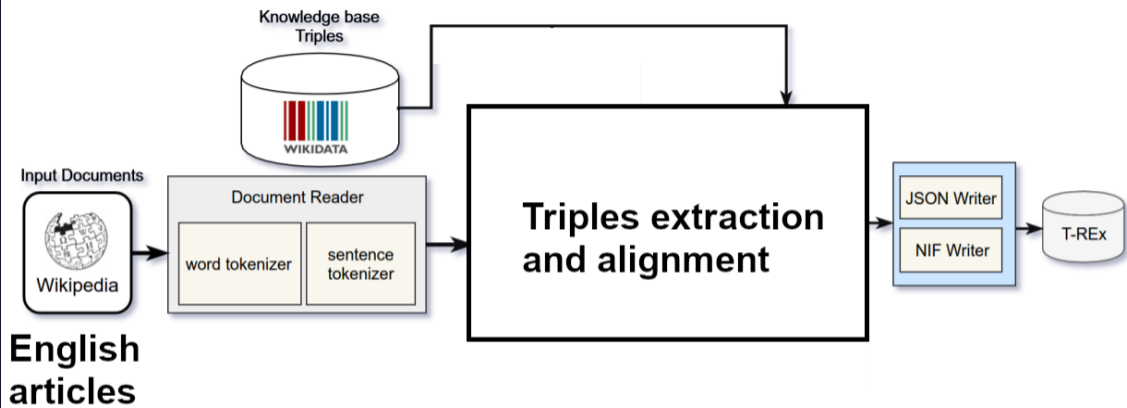








- Facts not on English Wikipedia?



**English  
articles**

- Facts not on English Wikipedia?
- An issue in a component of the pipeline?

# Impact of Bias on the results?



# Benchmark: mLAMA $\subset$ TRE-x with multilingual labels

---

Kassner, Nora, Dufter, Philipp, and Schütze, Hinrich. 2021. "Multilingual LAMA: Investigating Knowledge in Multilingual Pretrained Language Models."



# Benchmark:

## mLAMA $\subset$ TRE-x with multilingual labels

### Multilingual Triples:

- En: (Edinburgh [X], Country, Scotland)
- Ar: ([X] إدنبرة, بلد, اسكتلندا)

# Benchmark:

## mLAMA $\subset$ TRE-X with multilingual labels

### Multilingual Triples:

- En: (Edinburgh [X], Country, Scotland)
- Ar: ([X] إدنبرة, بلد, اسكتلندا)

### Manual Templates:

- [X] is located in ...
- ... تقع [X] في ...

# Benchmark:

## mLAMA $\subset$ TRE-x with multilingual labels

### Multilingual Triples:

- En: (Edinburgh [X], Country, Scotland)
- Ar: ([X] إدنبرة, بلد, اسكتلندا)

### Corresponding Prompts:

- Edinburgh is located in ...
- ... تقع إدنبرة في ...

# Benchmark:

## mLAMA $\subset$ TRE-x with multilingual labels

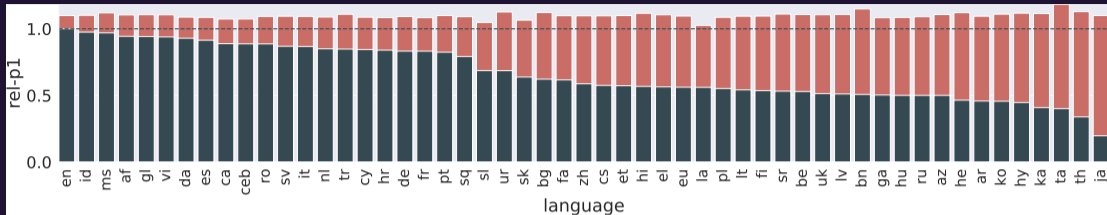
### Multilingual Triples:

- En: (Edinburgh [X], Country, Scotland)
- Ar: ([X] إدنبرة, بلد, اسكتلندا)

### Corresponding Prompts:

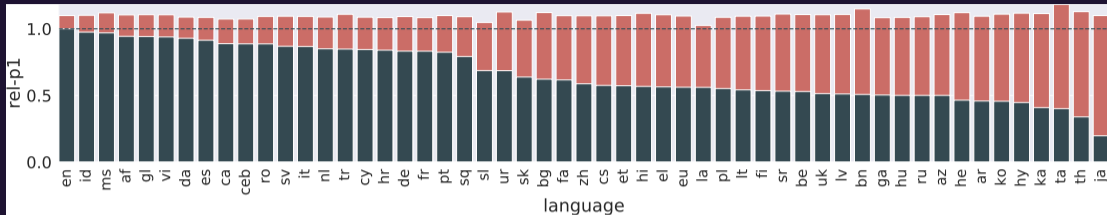
- Edinburgh is located in ...
- ... تقع إدنبرة في ...

## Model: mBERT



## For recalling facts from mBERT:

Ability<sub>Arabic prompts</sub>  $\approx \frac{1}{2}$  Ability<sub>English prompts</sub>

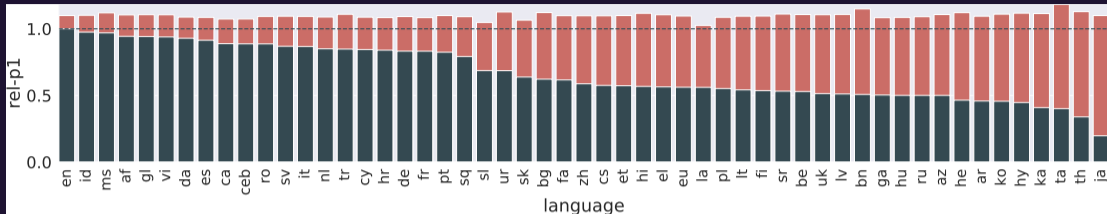


## For recalling facts from mBERT:

Ability<sub>Arabic prompts</sub>  $\approx \frac{1}{2}$  Ability<sub>English prompts</sub>

Possible explanations:

- mBERT's pretraining data
- Issues in prompts



## For recalling facts from mBERT:

Ability<sub>Arabic prompts</sub>  $\approx \frac{1}{2}$  Ability<sub>English prompts</sub>

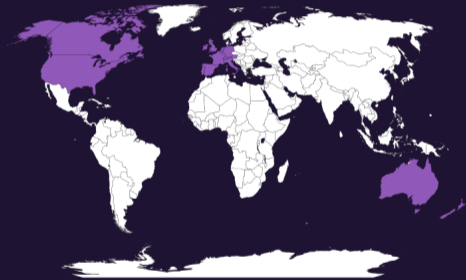
Possible explanations:

- mBERT's pretraining data
- Issues in prompts
- **Representation bias in the benchmark (mLAMA) ?**



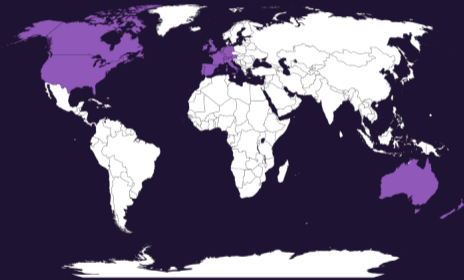
# Quantifying the Bias

- Identified **21 Western countries**.



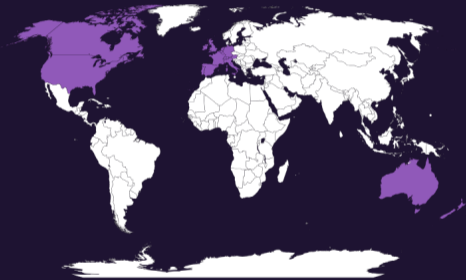
# Quantifying the Bias

- Identified **21 Western countries**.
- **26 relation predicates**.



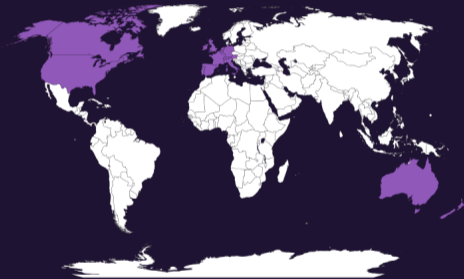
# Quantifying the Bias

- Identified **21 Western countries**.
- **26 relation predicates**.
- **TRE-x** (dump of facts),  
Benchmarks sampled from  
it: **LAMA, X-FACTR**



# Quantifying the Bias

- Identified **21 Western countries**.
- **26 relation predicates**.
- **TRE-x** (dump of facts),  
Benchmarks sampled from  
it: **LAMA, X-FACTR**
- Classify each triple as  
belonging to the **21  
Western countries** or not.



# Quantifying the Bias

≈ 1.5M triples (> 50%) of **T-REx** ∈ the **21 Western countries**

---

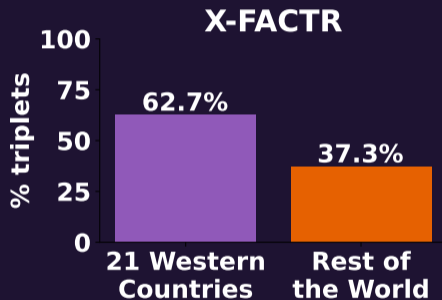
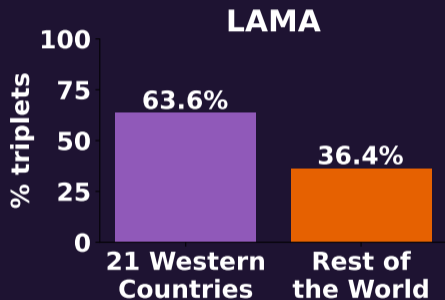
Petroni, Fabio et al. 2019. “Language Models as Knowledge Bases?”

Chaudhury, Subhjit et al. 2022. “X-FACTOR: A Cross-metric Evaluation of Factual Correctness in Abstractive Summarization.”



# Quantifying the Bias

≈ 1.5M triples (> 50%) of **T-REx** ∈ the **21 Western countries**



Petroni, Fabio et al. 2019. "Language Models as Knowledge Bases?"

Chaudhury, Subhjit et al. 2022. "X-FACTOR: A Cross-metric Evaluation of Factual Correctness in Abstractive Summarization."

# How to build **more diverse** **Multilingual** factual knowledge benchmarks?

# DLAMA's methodology

- 1 Identify cultures/regions of interest.
- 2 Curate facts related to these cultures/regions.



# (1) Identify cultures/regions of interest

- DLAMA-v1 has three splits of contrasting sets of facts;

# (1) Identify cultures/regions of interest

- DLAMA-v1 has three splits of contrasting sets of facts;
- spanning **20 Wikidata predicates**.

# (1) Identify cultures/regions of interest

- DLAMA-v1 has three splits of contrasting sets of facts;
- spanning **20 Wikidata predicates**.
- Contrast facts from **21 Western countries** to:

# (1) Identify cultures/regions of interest

- DLAMA-v1 has three splits of contrasting sets of facts;
- spanning **20 Wikidata predicates**.
- Contrast facts from **21 Western countries** to:
  - **22 Arab countries**
  - **12 South American countries**
  - **13 East-Asian & Southeast-Asian countries**

## (2) Curate culturally representative facts

**Predicate:** P27 (Country of Citizenship)

1 Query  
Wikidata

subj uri	obj uri	subj wikipedia url	article size
Q81324	Q30	.../Bret_Hart	201353
Q81328	Q30	.../Harrison_Ford	81422
Q65645	Q30	.../Matthias_Pintscher	6923
...	...	...	...

## (2) Curate culturally representative facts

**Predicate:** P27 (Country of Citizenship)

2 Sort

subj uri	obj uri	subj wikipedia url	article size
Q22686	Q30	.../Donald_Trump	417652
Q313381	Q30	.../Tom_Brady	408608
...	...	...	...

## (2) Curate culturally representative facts

**Predicate:** P27 (Country of Citizenship)

English	
subj label	obj label
Donald Trump	United States of America
Tom Brady	United States of America
...	...

Arabic	
subj label	obj label
دونالد ترامب	الولايات المتحدة
توم برايدي	الولايات المتحدة
...	...

### 3 Get parallel multilingual labels

- $\approx$  13K relation triples for each set of cultures.

**Arab facts****Western facts**

10,946 triples

13,589 triples

DLAMA-v1 (Arab-West)

**Asian facts****Western facts**

13,479 triples

13,588 triples

DLAMA-v1 (Asia-West)

**South American facts****Western facts**

13,071 triples

13,586 triples

DLAMA-v1 (South America-West)



- $\approx$  13K relation triples for each set of cultures.
- Parallel labels in **Arabic-English**, **Korean-English**, and **Spanish-English**.

### Arab facts

### Western facts

10,946 triples

13,589 triples

DLAMA-v1 (Arab-West)

### Asian facts

### Western facts

13,479 triples

13,588 triples

DLAMA-v1 (Asia-West)

### South American facts

### Western facts

13,071 triples

13,586 triples

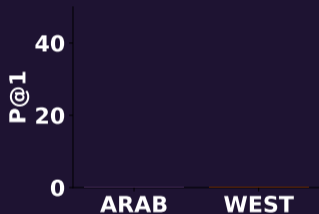
DLAMA-v1 (South America-West)

# Results!

# Results

## DLAMA-v1 (Arab-West)

Performance of **English BERT-base** on DLAMA-v1?

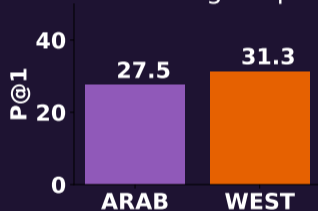


# Results

## DLAMA-v1 (Arab-West)

Performance of **English BERT-base** on DLAMA-v1?

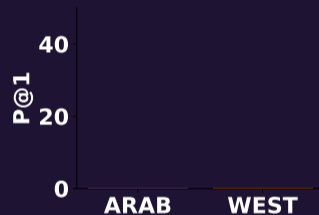
BERT-base - English prompts



# Results

## DLAMA-v1 (Arab-West)

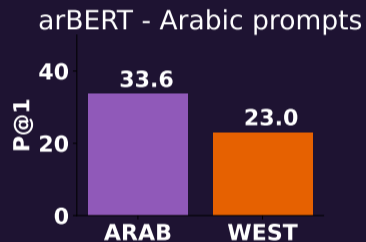
Performance of **Arabic arBERT** on DLAMA-v1?



# Results

## DLAMA-v1 (Arab-West)

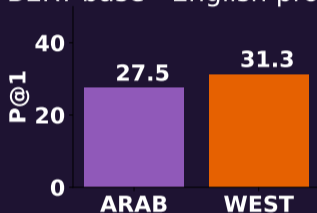
Performance of **Arabic arBERT** on DLAMA-v1?



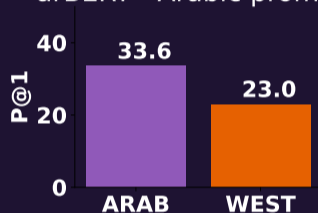
# Results

## DLAMA-v1 (Arab-West)

BERT-base - English prompts



arBERT - Arabic prompts

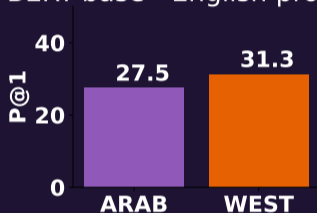


- Prompts in a language better for related facts;

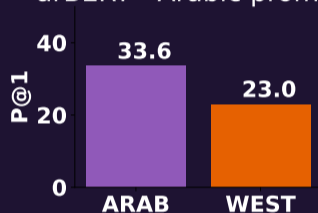
# Results

## DLAMA-v1 (Arab-West)

BERT-base - English prompts



arBERT - Arabic prompts



- Prompts in a language better for related facts;
- similar behaviours for the other two sets of facts

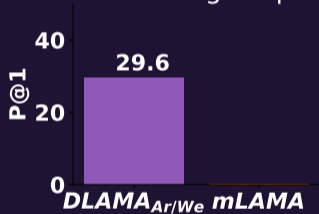


# Results

## DLAMA-v1 (Arab-West)

Performance of **English BERT-base** on mLAMA vs DLAMA-v1?

BERT-base - English prompts

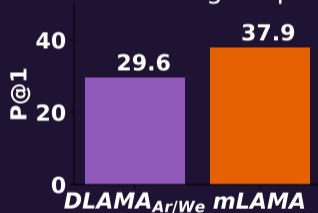


# Results

## DLAMA-v1 (Arab-West)

Performance of **English BERT-base** on mLAMA vs DLAMA-v1?

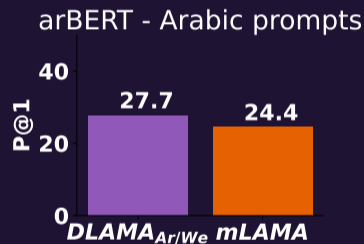
BERT-base - English prompts



# Results

## DLAMA-v1 (Arab-West)

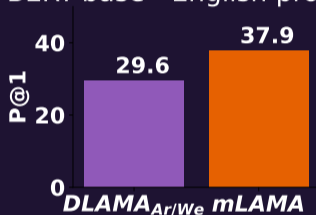
Performance of **Arabic arBERT** on mLAMA vs DLAMA-v1



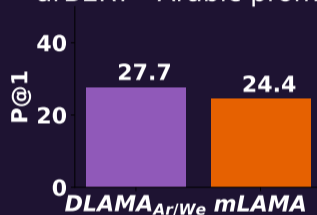
# Results

## DLAMA-v1 (Arab-West)

BERT-base - English prompts



arBERT - Arabic prompts



**Biased benchmarks** ⇒ **Biased evaluation!**

# Thanks!

 @Amrkeleg



# Thanks!

 @Amrkeleg



## Summary:

- Translating English benchmarks introduces a **cultural bias**.
- We introduce **DLAMA-v1**, a benchmark with facts from 3 **contrasting** sets of cultures.
- Contrasting cultures provides better **Model Analysis**.

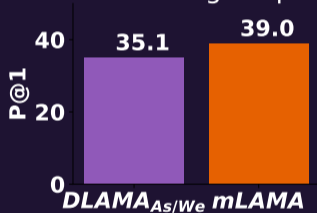
# GPT3.5 turbo QA performance

Relation	Arabic prompts Accuracy		English prompts Accuracy	
	Arab facts	West facts	Arab facts	West facts
P30 (Continent)	63.6	<b>89.5*</b>	<b>100.0*</b>	89.5
P36 (Capital)	<b>81.8*</b>	63.2	<b>95.5*</b>	94.7
P37 (Official language)	<b>100.0*</b>	89.5	<b>100.0*</b>	<b>100.0*</b>
P47 (Shares border with)	<b>100.0*</b>	<b>100.0*</b>	<b>95.5*</b>	89.5
P190 (Sister city)	<b>6.0*</b>	5.6	3.0	<b>33.1*</b>
P530 (Diplomatic relation)	63.6	<b>68.4*</b>	50.0	<b>84.2*</b>
P1376 (Capital of)	87.5	<b>88.5*</b>	<b>100.0*</b>	92.3

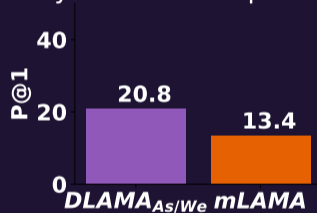
# Results

## DLAMA-v1 (Asia-West)

BERT-base - English prompts



KyKim - Korean prompts

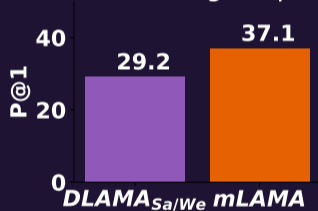




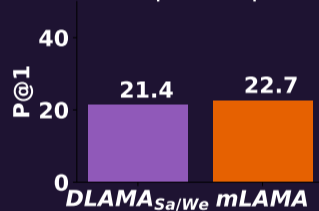
# Results

## DLAMA-v1 (South America-West)

BERT-base - English prompts



BETO - Spanish prompts



## 21 Western Countries:

The Inner Circle, Western European and South Western European countries:

Andorra, Austria, Belgium, France, Germany, Ireland, Italy, Liechtenstein, Luxembourg, Monaco, Netherlands, Portugal, San Marino, Spain, Switzerland, the United Kingdom, in addition to Canada, the United States of America, Australia, and New Zealand.

## **12 South American countries:**

Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, Guyana, Paraguay, Peru, Suriname, Uruguay, Venezuela

## **13 East Asian, and Southeast Asian countries:**

China, Indonesia, Japan, Malaysia, Mongolia, Myanmar, North Korea, Philippines, Singapore, South Korea, Taiwan, Thailand, Vietnam

# Why contrastive sets of facts?

# Benefits of contrasting results

**Model:** English BERT-Base.

**Prompt for P495 (Country of Origin) :** [X] was created in [Y] .

# Benefits of contrasting results

**Model:** English BERT-Base.

**Prompt for P495 (Country of Origin) :** [X] was created in [Y] .

Performance<sub>Asian facts</sub>  $\gg$  Performance<sub>Western facts</sub>

# Benefits of contrasting results

**Model:** English BERT-Base.

**Prompt for P495 (Country of Origin) :** [X] was created in [Y] .

Performance<sub>Asian facts</sub>  $\gg$  Performance<sub>Western facts</sub>

 **WHY?!**

# Benefits of contrasting results

**Model:** English BERT-Base.

**Prompt for P495 (Country of Origin) :** [X] was created in [Y] .

## Model Predictions for Asian facts



# Benefits of contrasting results

**Model:** English BERT-Base.

**Prompt for P495 (Country of Origin) :** [X] was created in [Y] .

## Model Predictions for Asian facts

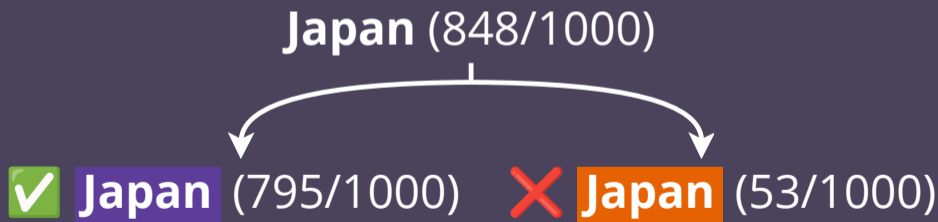
Japan (848/1000)

# Benefits of contrasting results

**Model:** English BERT-Base.

**Prompt for P495 (Country of Origin) :** [X] was created in [Y] .

## Model Predictions for Asian facts



# Benefits of contrasting results

**Model:** English BERT-Base.

**Prompt for P495 (Country of Origin) :** [X] was created in [Y] .

## Model Predictions for Western facts

# Benefits of contrasting results

**Model:** English BERT-Base.

**Prompt for P495 (Country of Origin) :** [X] was created in [Y] .

## Model Predictions for Western facts

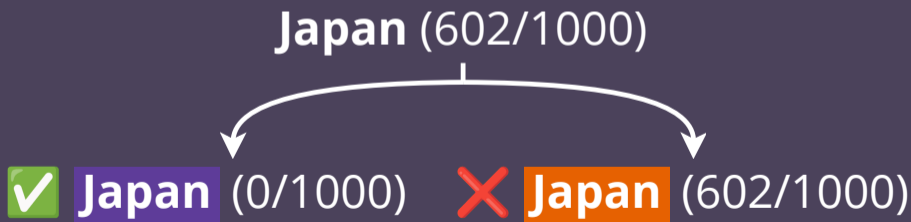
Japan (602/1000)

# Benefits of contrasting results

**Model:** English BERT-Base.

**Prompt for P495 (Country of Origin) :** [X] was created in [Y] .

## Model Predictions for Western facts



# Benefits of contrasting results

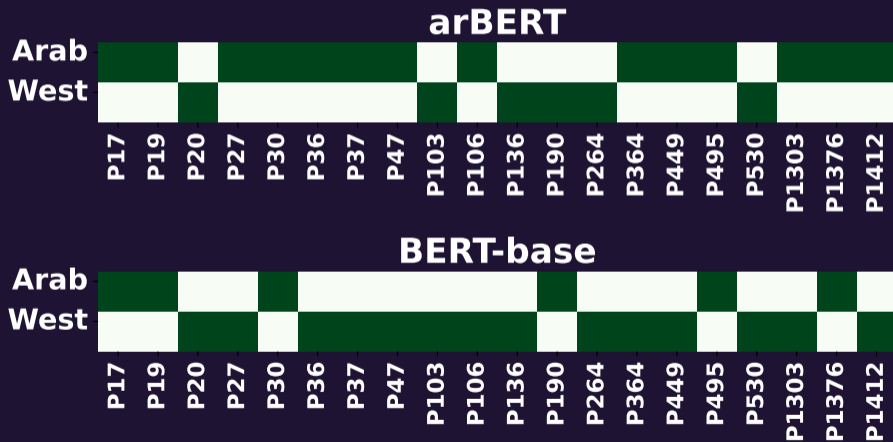
**Model:** English BERT-Base.

**Prompt for P495 (Country of Origin) :** [X] was created in [Y] .

Prompt Bias toward **Japan!**

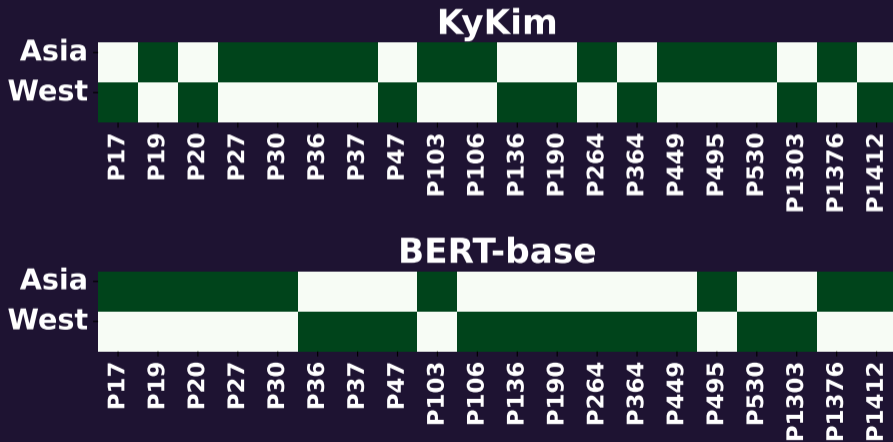
# Heatmap of individual predicates

## DLAMA-v1 (Arab-West)



# Heatmap of individual predicates

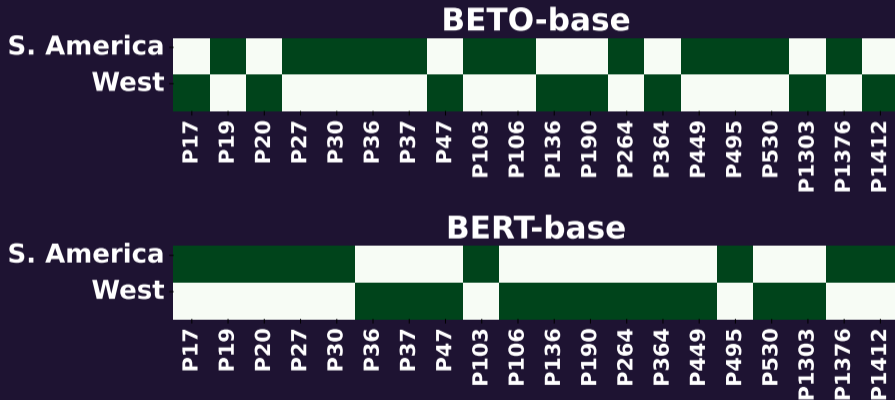
## DLAMA-v1 (Asia-West)






# Heatmap of individual predicates

## DLAMA-v1 (South America-West)




# References I

-  Chaudhury, Subhajit et al. (Dec. 2022). “X-FACTOR: A Cross-metric Evaluation of Factual Correctness in Abstractive Summarization.” In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 7100–7110. URL: <https://aclanthology.org/2022.emnlp-main.478>.


# References II

-  Elshahar, Hady et al. (May 2018). “T-REx: A Large Scale Alignment of Natural Language with Knowledge Base Triples.” In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA). URL: <https://aclanthology.org/L18-1544>.

# References III

-  Kassner, Nora, Philipp Dufter, and Hinrich Schütze (Apr. 2021). “Multilingual LAMA: Investigating Knowledge in Multilingual Pretrained Language Models.” In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online: Association for Computational Linguistics, pp. 3250–3258. DOI: [10.18653/v1/2021.eacl-main.284](https://doi.org/10.18653/v1/2021.eacl-main.284). URL: <https://aclanthology.org/2021.eacl-main.284>.

# References IV

-  Petroni, Fabio et al. (Nov. 2019). “Language Models as Knowledge Bases?” In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, pp. 2463–2473. DOI: 10.18653/v1/D19-1250. URL: <https://aclanthology.org/D19-1250>.