1)**Cultural Bias** in Knowledge Benchmarks.
2)**DLAMA-v1**: Culturally Diverse Benchmark.
3)**Contrastive** Sets of Facts Enrich Analysis.

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

Amr Keleg, Walid Magdy

Factual Knowledge of mBERT

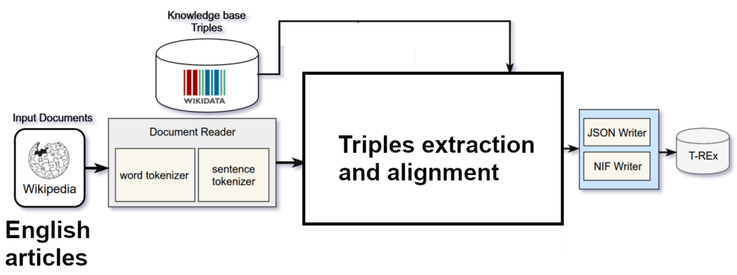
- Using prompts from **mLAMA** benchmark

$$\text{Acc}_{\text{Arabic prompts}} < \frac{1}{2}\,\text{Acc}_{\text{English prompts}}$$

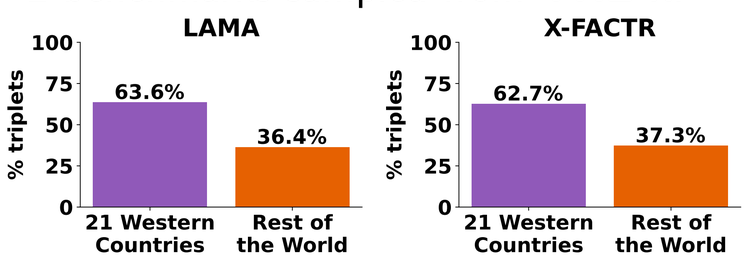
- Is **mLAMA** culturally representative?

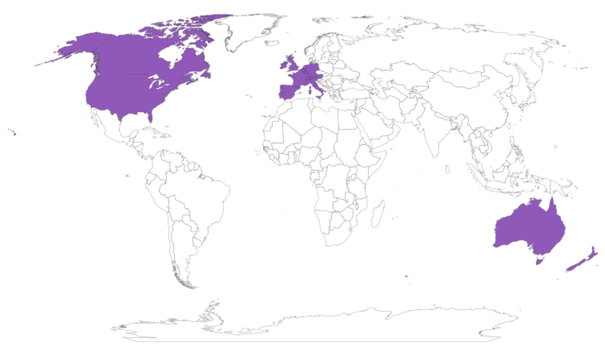
Bias in Knowledge Benchmarks

- **TRE-x** creation steps **(problems?)**:



- 2 benchmarks sampled from **TRE-x**:





Probing Setup

- Staring from Wikidata triplets:

  Subject: Edward I of England [X]

  Predicate: religion or worldview [P140]

  Object: Christianity [Y]

- Manual templates to form prompts:

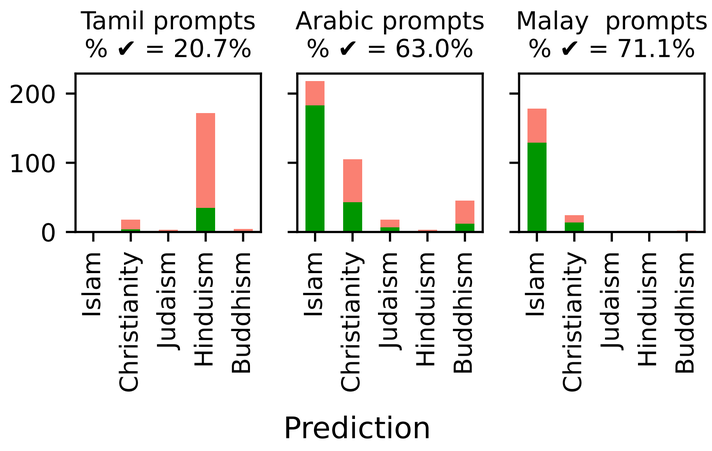
  Edward I of England is affiliated with ...

  (a) Islam **(b) Christianity** (c) Judaism (d) Hinduism (e) Buddhism ...

- **Choices:** objects **[Y]** for the predicate.

Impact of Bias on Analysis

- **Benchmark**: mLAMA - P140
- Tamil, Arabic, Malay prompts
- **Islam** is the most common object.



- ⚠️ # prompts differs across langauges.
- Malay prompts > Arabic prompts?

The DLAMA Framework

1. Set parameters

Input

Query parameters

- **Subject:** Person
- **Object:** Country
- **Predicate:** P27 (Country of Citizenship)
- **Countries of interest:** USA
- **Languages of labels:** English, Arabic

SPARQL query → Wikidata

2. Query Wikidata triplets

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

3. Sort by article size

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

4. Query multilingual labels

English		Arabic	
subj label	obj label	subj label	obj label
Donald Trump	United States of America	دونالد ترامب	الولايات المتحدة
Tom Brady	United States of America	توم برايدي	الولايات المتحدة
...	...	...	...

The DLAMA-v1 Benchmark

- > **78K** triplets from **20 predicates**
- 3 **bilingual** contrasting sets of facts:
  - **Western(en)** ⚔️ **Arab(ar)**
  - **Western(en)** ⚔️ **South American(es)**
  - **Western(en)** ⚔️ **East-Asian(ko)**

Results

DLAMA-v1 (Arab-West)

Prompt Lang.	Model name	$P_{@1}$ Arab N=10946	$P_{@1}$ West N=13589
Arabic	mBERT-base	13.7	**15.1***
	arBERT	**33.6***	23.0
English	mBERT-base	21.2	**37.7***
	BERT-base	27.5	**31.3***

- arBERT$_{\text{(Arab)}}$ on par with BERT-base$_{\text{(West)}}$

DLAMA-v1 (South America-West)

Prompt Lang.	Model name	$P_{@1}$ S. America N=13071	$P_{@1}$ West N=13586
Spanish	mBERT-base	25.4	**33.8***
	BETO	16.0	**26.5***
English	mBERT-base	27.0	**37.6***
	BERT-base	26.9	**31.3***

- Models not capturing S. American facts?

DLAMA-v1 (Asia-West)

Prompt Lang.	Model name	$P_{@1}$ Asia N=13479	$P_{@1}$ West N=13588
Korean	mBERT-base	16.4	**28.5***
	KyKim	**22.1***	19.5
English	mBERT-base	33.0	**39.9***
	BERT-base	**38.3***	31.9

- BERT-base$_{\text{(Asia)}}$ > BERT-base$_{\text{(West)}}$ ⬇️

Benefits of Contrastive Benchmarks

Model: English BERT-base

Prompt [P495]: [X] was created in [Y] .

English prompts of Asian facts

Most Common ✅ prediction (% of all predictions)	Most Common ❌ prediction (% of all predictions)
Japan (79.5%)	Japan (5.3%)

English prompts of Western facts

Most Common ✅ prediction (% of all predictions)	Most Common ❌ prediction (% of all predictions)
France (1.5%)	Japan (60.2%)

- **Prompt bias** toward **Japan**?

THE UNIVERSITY of EDINBURGH informatics
ilcc Institute for Language, Cognition and Computation
NLP UKRI CENTRE FOR DOCTORAL TRAINING