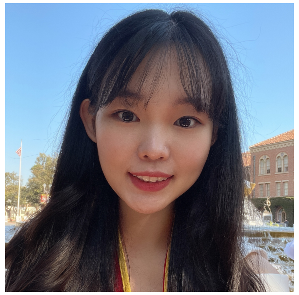
ACL-IJCNLP 2021

http://inklab.usc.edu/XCSR/

Common Sense Beyond English: Evaluating and Improving Multilingual Language Models for Commonsense Reasoning



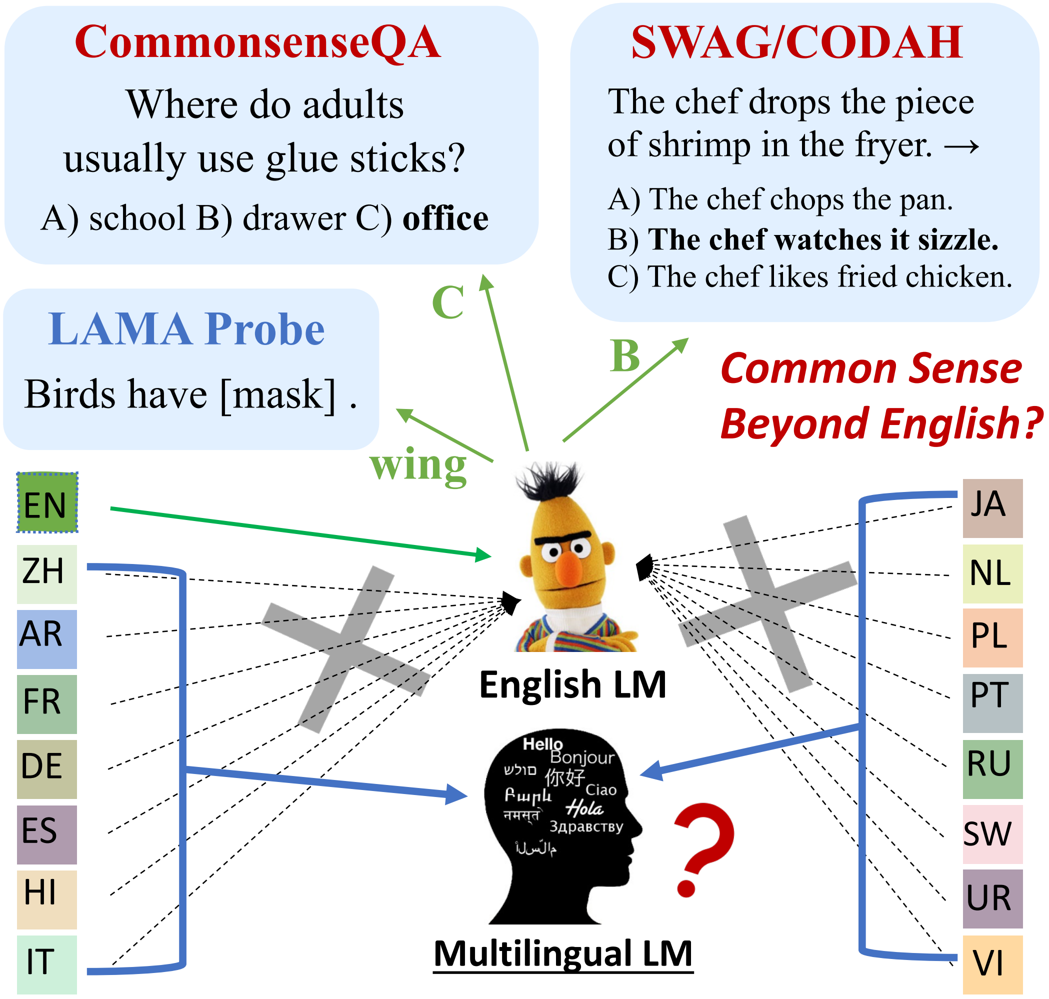
Bill Yuchen Lin **Seyeon Lee** **Xiaoyang Qiao** **Xiang Ren**

{yuchen.lin, seyeonle, xiaoyanq, xiangren}@usc.edu

Department of Computer Science and Information Sciences Institute,
University of Southern California

USC

Common Sense Beyond English and En-LMs

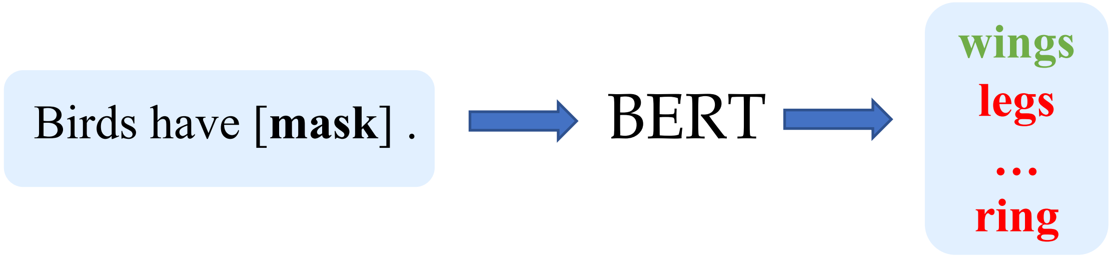

CommonsenseQA
Where do adults usually use glue sticks?
A) school B) drawer C) office
SWAG/CODAH
The chef drops the piece of shrimp in the fryer. →
A) The chef chops the pan.
B) The chef watches it sizzle.
C) The chef likes fried chicken.
LAMA Probe
Birds have [mask] .
C
B
wing
Common Sense Beyond English?
EN
ZH
AR
FR
DE
ES
HI
IT
JA
NL
PL
PT
RU
SW
UR
VI
English LM
Hello
Bonjour
你好
Ciao
Hola
Здравству
Multilingual LM
?

- Common Sense research has been limited to English and En-LMs.
  - **Probing** (e.g., LAMA)
  - **Reasoning** (e.g., CSQA, CODAH)

- What about other languages and multilingual LMs (ML-LMs)?
  - Common Sense Beyond En and En-LMs.

Background: LAMA Probing and Its Limitations

- LAMA probing (EMNLP'19) as a **zero-shot token-ranking task.**



Birds have [**mask**] . → BERT →

wings
legs
…
ring

- Limitations
  - Only token-level predictions.
    - **Multi-token words** can be very common in other languages, e.g., Chinese.
    - **Sentence-level common sense** is not explicitly tested.
  - Evaluation metric.
    - Different langs have totally **different vocabulary**.
    - **False negative** is problematic because of corpus bias.

Mickey Probe: Multilingual Commonsense Knowledge Probing

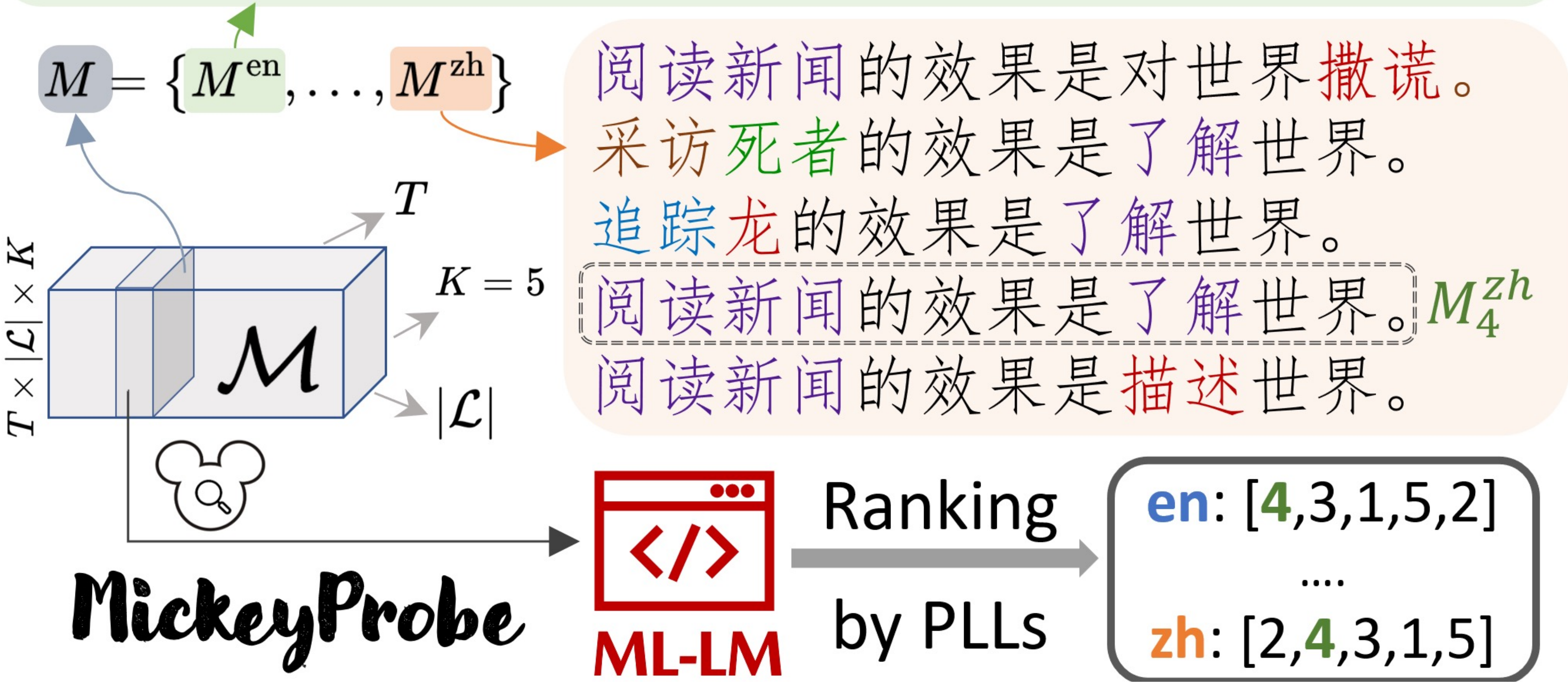
The effect of reading the news is lying about the world.

… of interviewing the deceased is learning about the world.

… of tracking the dragon is learning about the world.

… of reading the news is learning about the world. M_4^{en}

… of reading the news is saying about the world.



阅读新闻的效果是对世界撒谎。

采访死者的效果是了解世界。

追踪龙的效果是了解世界。

阅读新闻的效果是了解世界。M_4^{zh}

阅读新闻的效果是描述世界。

en: [4,3,1,5,2]

….

zh: [2,4,3,1,5]

MickeyProbe as Sent-Ranking

Input:

- A set of (K=5) *statements*.
- There are $|\mathcal{L}|$ different *versions* in testing languages.

$$\mathcal{L} = \{en, fr, ru, zh, \ldots\}$$

$$\forall M \in \mathcal{M},\ \forall (l_x, l_y) \in \mathcal{L}^2,\ \forall i \in \mathbb{N}_{\leq K},\ M_i^{l_x} \bowtie M_i^{l_y}.$$

Output:

- A sorted list of sentences in each lang's version.
- The **top-1** should have highest commonsense plausibility.

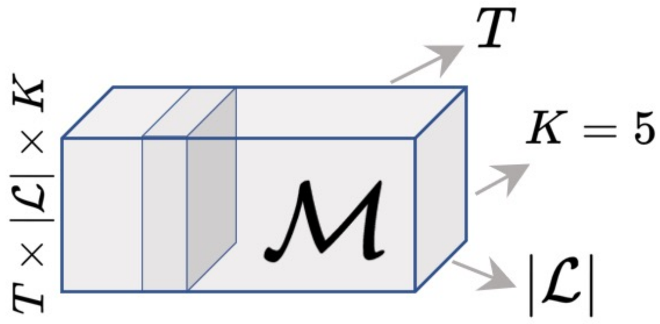
Probing w/ PLLs (Salazar; ACL'20):

$$\log P(\boldsymbol{s}) = \operatorname{PLL}(\boldsymbol{s}) := \sum_{i=1}^{|\boldsymbol{s}|} \log P\left(w_i \mid \boldsymbol{s}_{\backslash i}\right)$$

Constructing the Corpus for Mickey Probe

- OMCS corpus: a collection of commonsense assertions in English.
  - Step 1: generate distractors via masked-attacking.
  - Step 2: machine translation with MarianMT for En-X.
  - Step 3: back-translation (En-X-EN) and filter with SentenceBERT.

More implementation details are in our paper.



$T = 10.2\text{k}$, $K = 5$, $|\mathcal{L}| = 11$

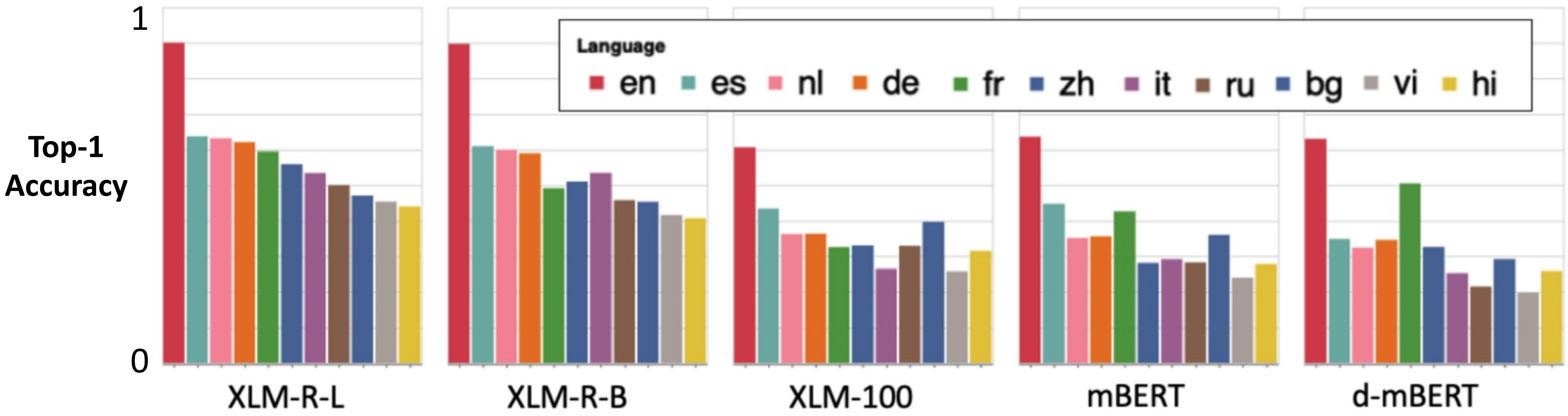
T = # probes.

K = # options (K-1 distractors)

|*L*| = # languages.

561k sentences in total.

Mickey Probing Results



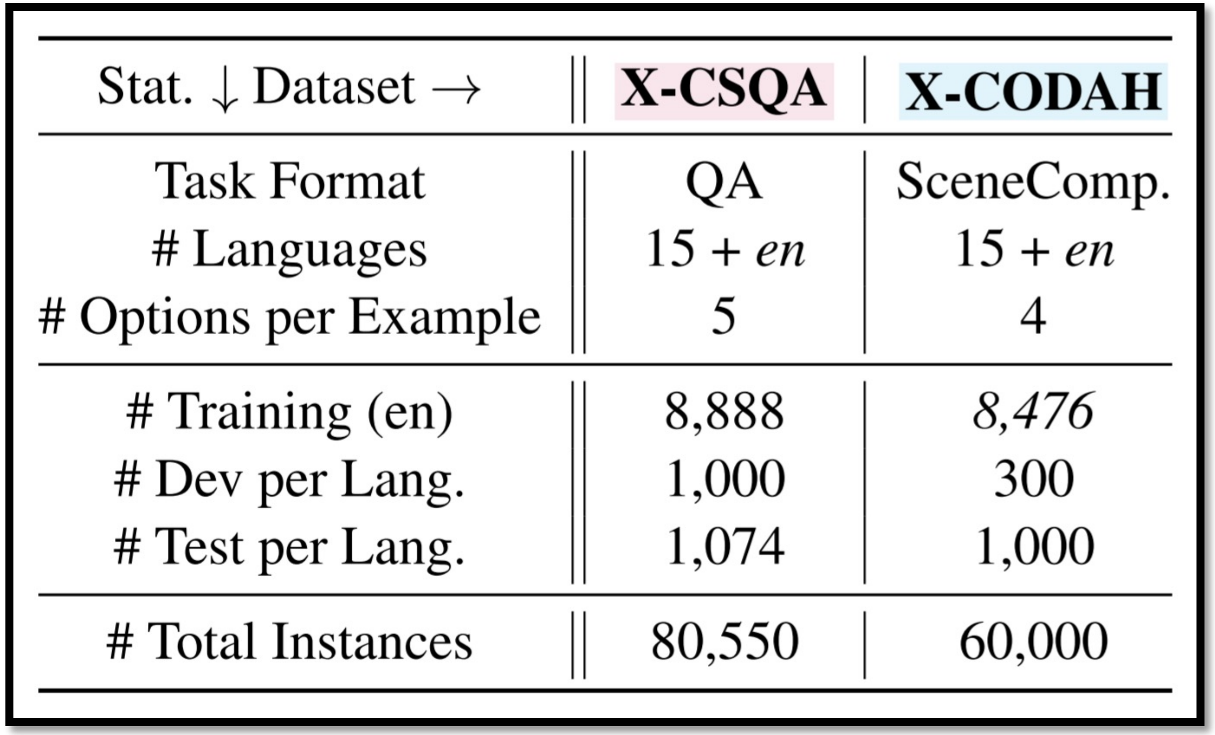
Q1*: Do different ML-LMs have similar language preferences?* ***No***

Q2*: Does length influence PLL-based ranking?* ***Not much.*** *See the "Shortest" row.*

Corpus size and translation quality are not ***major factors*** *in lang preference of ML-LMs.*

Further implications: *common sense can be a bridge for the alignment within ML-LMs.*

X-CSR Datasets and Evaluation



Stat. ↓ Dataset →	X-CSQA	X-CODAH
Task Format	QA	SceneComp.
# Languages	15 + *en*	15 + *en*
# Options per Example	5	4
# Training (en)	8,888	*8,476*
# Dev per Lang.	1,000	300
# Test per Lang.	1,074	1,000
# Total Instances	80,550	60,000

Zero-Shot Cross-Lingual Transfer

- Initialize pre-trained ML-LMs
- Fine-tune on English Training data.
- Validate and test on all languages.

Shared setup (and langs) w/ XNLI.

{en, zh, de, es, fr, it, jap, nl, pl, pt, ru, ar, vi, hi, **sw, ur**}

↓

low-resource

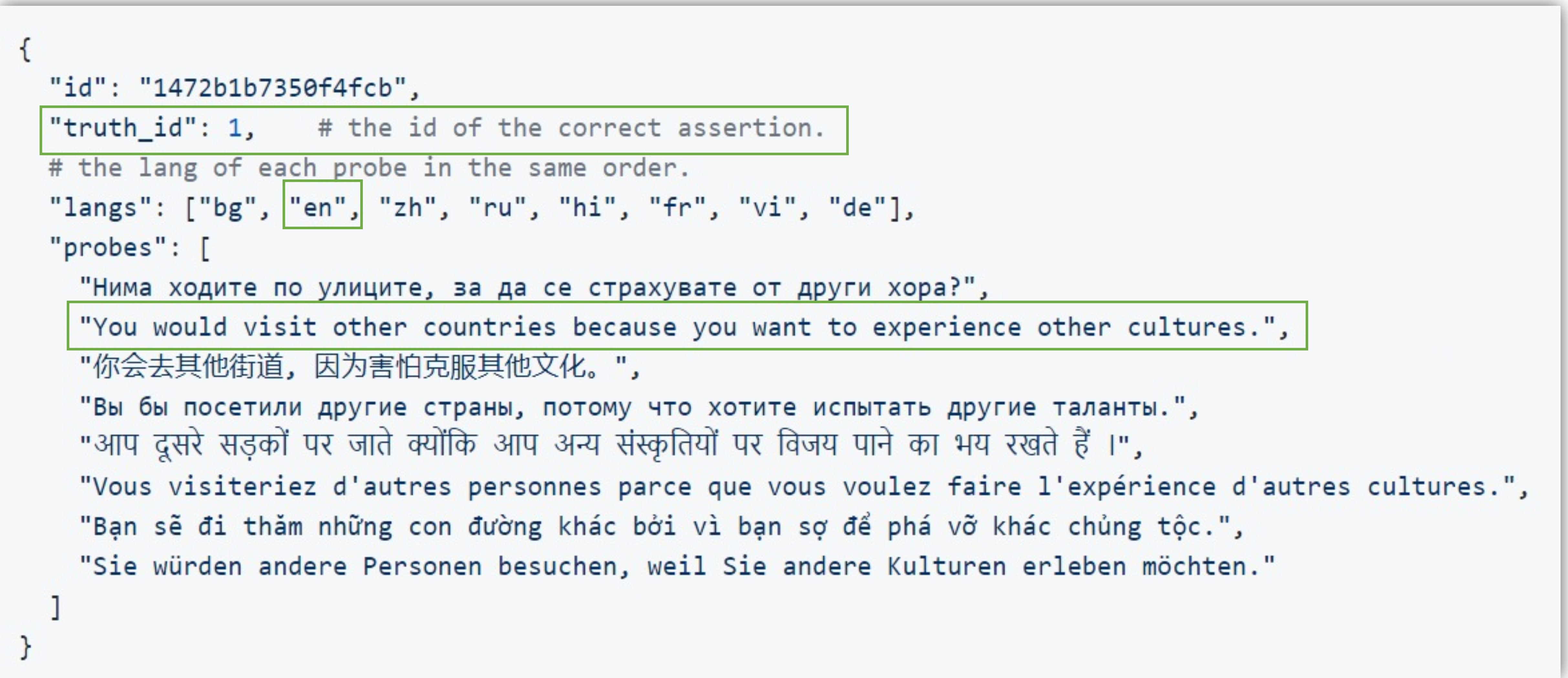
MCP: Multilingual Contrastive Pre-training of ML-LMs.

- **Background**: most ML-LMs only have token-level objectives during pre-training as follows:
  - **Masked-LM** (masked-word-prediction inside a single language)
  - **Translation-LM** (a cross-lingual version in a pair of parallel sentences.)
- **Motivation**:
  - The key to NLU tasks is strong sentence representation ([CLS]).
  - However, it is not well pre-trained.
  - Can we design a multilingual, sentence-level pre-training task?

MCP: Multilingual Contrastive Pre-training of ML-LMs.

- **Goal**: reuse the Mickey corpus to create a multilingual task.
- **Start**: a **MickeyProbe** instance.
  - Step 1): we randomly pick a lang as the anchor lang.
  - Step 2): use the correct assertion in the anchor lang as the target sentence.
  - Step 3): randomly select distractors from other languages.
- **End**: a **MCP** task instance.
  - **MCP input:** a set of sampled sentences in different languages, where only one is a commonsense assertion and others are distracotrs.
  - **MCP output:** the label to the correct commonsense assertion (via [CLS] representation).

An example of MCP instance.

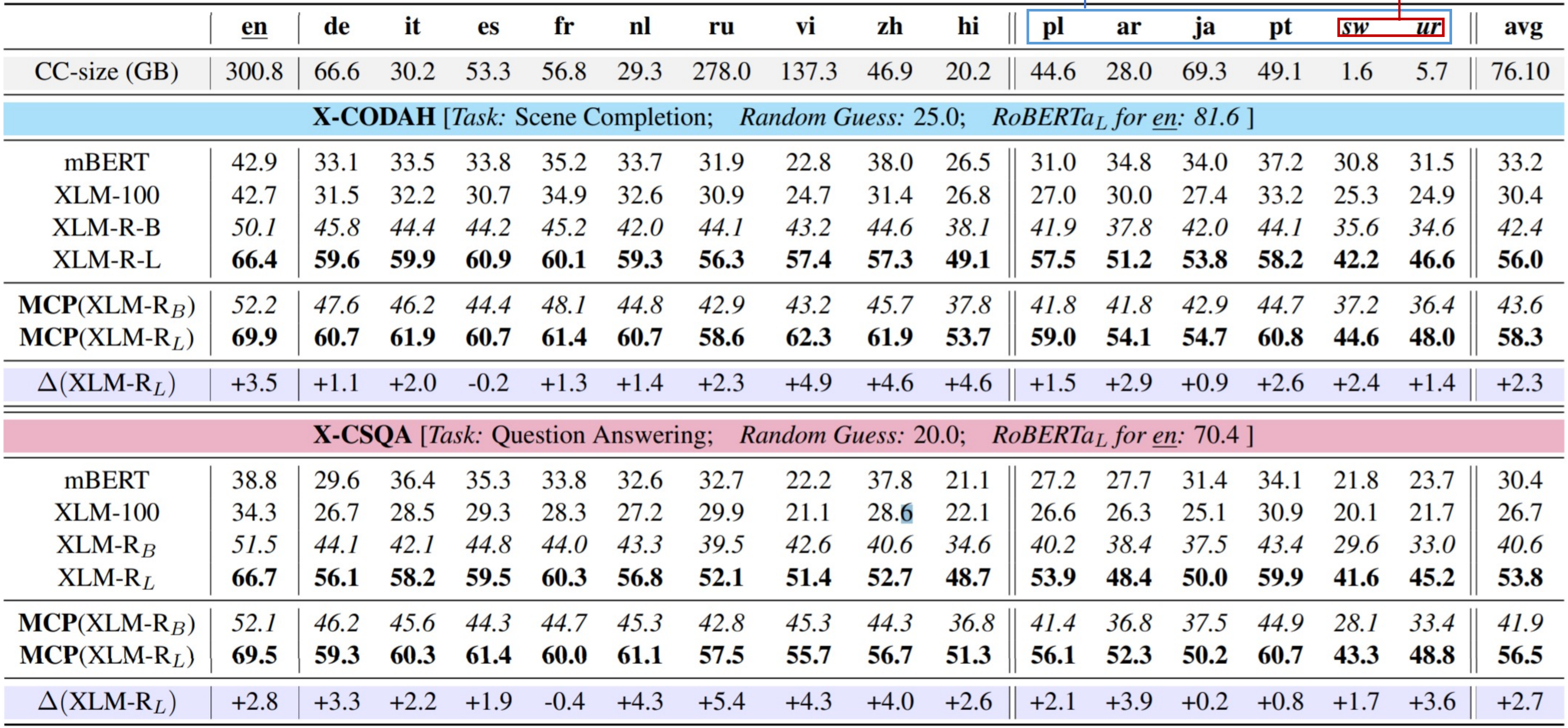


```
{
  "id": "1472b1b7350f4fcb",
  "truth_id": 1,    # the id of the correct assertion.
  # the lang of each probe in the same order.
  "langs": ["bg", "en", "zh", "ru", "hi", "fr", "vi", "de"],
  "probes": [
    "Нима ходите по улиците, за да се страхувате от други хора?",
    "You would visit other countries because you want to experience other cultures.",
    "你会去其他街道，因为害怕克服其他文化。",
    "Вы бы посетили другие страны, потому что хотите испытать другие таланты.",
    "आप दूसरे सड़कों पर जाते क्योंकि आप अन्य संस्कृतियों पर विजय पाने का भय रखते हैं ।",
    "Vous visiteriez d'autres personnes parce que vous voulez faire l'expérience d'autres cultures.",
    "Bạn sẽ đi thăm những con đường khác bởi vì bạn sợ để phá vỡ khác chủng tộc.",
    "Sie würden andere Personen besuchen, weil Sie andere Kulturen erleben möchten."
  ]
}
```

X-CSR Experimental Results

Unseen langs in MCP.

Low-resource.



	en	de	it	es	fr	nl	ru	vi	zh	hi	pl	ar	ja	pt	*sw*	*ur*	avg
CC-size (GB)	300.8	66.6	30.2	53.3	56.8	29.3	278.0	137.3	46.9	20.2	44.6	28.0	69.3	49.1	1.6	5.7	76.10
X-CODAH [*Task:* Scene Completion; *Random Guess:* 25.0; *RoBERTa$_L$ for en: 81.6*]																	
mBERT	42.9	33.1	33.5	33.8	35.2	33.7	31.9	22.8	38.0	26.5	31.0	34.8	34.0	37.2	30.8	31.5	33.2
XLM-100	42.7	31.5	32.2	30.7	34.9	32.6	30.9	24.7	31.4	26.8	27.0	30.0	27.4	33.2	25.3	24.9	30.4
XLM-R-B	*50.1*	*45.8*	*44.4*	*44.2*	*45.2*	*42.0*	*44.1*	*43.2*	*44.6*	*38.1*	*41.9*	*37.8*	*42.0*	*44.1*	*35.6*	*34.6*	*42.4*
XLM-R-L	**66.4**	**59.6**	**59.9**	**60.9**	**60.1**	**59.3**	**56.3**	**57.4**	**57.3**	**49.1**	**57.5**	**51.2**	**53.8**	**58.2**	**42.2**	**46.6**	**56.0**
MCP(XLM-R$_B$)	*52.2*	*47.6*	*46.2*	*44.4*	*48.1*	*44.8*	*42.9*	*43.2*	*45.7*	*37.8*	*41.8*	*41.8*	*42.9*	*44.7*	*37.2*	*36.4*	*43.6*
MCP(XLM-R$_L$)	**69.9**	**60.7**	**61.9**	**60.7**	**61.4**	**60.7**	**58.6**	**62.3**	**61.9**	**53.7**	**59.0**	**54.1**	**54.7**	**60.8**	**44.6**	**48.0**	**58.3**
Δ(XLM-R$_L$)	+3.5	+1.1	+2.0	-0.2	+1.3	+1.4	+2.3	+4.9	+4.6	+4.6	+1.5	+2.9	+0.9	+2.6	+2.4	+1.4	+2.3
X-CSQA [*Task:* Question Answering; *Random Guess:* 20.0; *RoBERTa$_L$ for en:* 70.4]																	
mBERT	38.8	29.6	36.4	35.3	33.8	32.6	32.7	22.2	37.8	21.1	27.2	27.7	31.4	34.1	21.8	23.7	30.4
XLM-100	34.3	26.7	28.5	29.3	28.3	27.2	29.9	21.1	28.6	22.1	26.6	26.3	25.1	30.9	20.1	21.7	26.7
XLM-R$_B$	*51.5*	*44.1*	*42.1*	*44.8*	*44.0*	*43.3*	*39.5*	*42.6*	*40.6*	*34.6*	*40.2*	*38.4*	*37.5*	*43.4*	*29.6*	*33.0*	*40.6*
XLM-R$_L$	**66.7**	**56.1**	**58.2**	**59.5**	**60.3**	**56.8**	**52.1**	**51.4**	**52.7**	**48.7**	**53.9**	**48.4**	**50.0**	**59.9**	**41.6**	**45.2**	**53.8**
MCP(XLM-R$_B$)	*52.1*	*46.2*	*45.6*	*44.3*	*44.7*	*45.3*	*42.8*	*45.3*	*44.3*	*36.8*	*41.4*	*36.8*	*37.5*	*44.9*	*28.1*	*33.4*	*41.9*
MCP(XLM-R$_L$)	**69.5**	**59.3**	**60.3**	**61.4**	**60.0**	**61.1**	**57.5**	**55.7**	**56.7**	**51.3**	**56.1**	**52.3**	**50.2**	**60.7**	**43.3**	**48.8**	**56.5**
Δ(XLM-R$_L$)	+2.8	+3.3	+2.2	+1.9	-0.4	+4.3	+5.4	+4.3	+4.0	+2.6	+2.1	+3.9	+0.2	+0.8	+1.7	+3.6	+2.7

Take-home Messages

- **MickeyProbe** task can analyze commonsense knowledge of ML-LMs.
  - Multilingual, sentence-level extension of LAMA probing.
  - It can compare performance across ML-LMs & across languages.
  - **MickeyCorpus** is a large multilingual corpus for studying common sense beyond English and English LMs. (561k sentences, in 11 languages.)
- **X-CSR** is a benchmark with two datasets (X-CSQA & X-CODAH) for zero-shot x-l transfer.
- **MCP** is a method to improve sentence rep. via multilingual contrastive pre-training.
  - Reusing the Mickey corpus for a multilingual sentence selection task.
  - Largely improve the X-CSR performance, even for unseen langs and low-resource langs.

Thank you very much for listening!

Please check our paper and project website for more info.

http://inklab.usc.edu/XCSR/