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

Abstract

Commonsense reasoning research has so far been limited to English. We aim to evaluate and improve popular multilingual language models (ML-LMs) to help advance commonsense reasoning (CSR) beyond English. We collect the Mickey corpus, consisting of 561k sentences in 11 different languages, which can be used for *analyzing* and *improving* ML-LMs. We propose Mickey Probe, a *language-agnostic* probing task for fairly evaluating the common sense of popular ML-LMs across different languages. In addition, we also create two new datasets, X-CSQA and X-CODAH, by translating their English versions to 15 other languages, so that we can evaluate popular ML-LMs for cross-lingual commonsense reasoning. To improve the performance beyond English, we propose a simple yet effective method — multilingual contrastive pre-training (MCP). It significantly enhances sentence representations, yielding a large performance gain on both benchmarks (e.g., +2.7% accuracy for X-CSQA over XLM-R$_L$)[1].

1 Introduction

Understanding natural language relies heavily on commonsense reasoning (CSR), which is the process of making inferences with commonsense knowledge. Commonsense knowledge is the set of general facts that reflect our natural understanding of the physical world and human behavior, which are usually seen as an implicit background when people communicate with each other using languages. It is thus of vital importance to evaluate and improve the commonsense reasoning capability of language models (LMs), towards building general natural language understanding (NLU) systems (Davis and Marcus, 2015).

[1]We release our code and data at the project website: https://inklab.usc.edu/XCSR/.

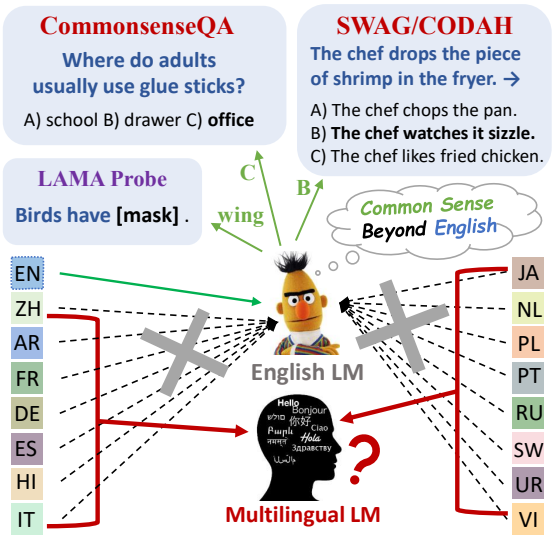

Figure 1: Commonsense reasoning is well-studied with benchmarks and LMs in English. Can we advance commonsense reasoning beyond English?

Many recent benchmark datasets and probing methods have been proposed to evaluate machine common sense. As shown in Figure 1, the LAMA probe (Petroni et al., 2019) is for analyzing LMs' *zero-shot* commonsense recalling ability; CommonsenseQA (CSQA) (Talmor et al., 2019) is instead a multiple-choice QA task that needs fine-tuning; CODAH (Chen et al., 2019) and SWAG (Zellers et al., 2018) focus on the ability to complete the most plausible scenes. However, all these works have been limited only to *English*. Consequently, follow-up analysis and reasoning methods developed (Lin et al., 2019; Feng et al., 2020; Lin et al., 2020) also focus only on *English* LMs like BERT (Devlin et al., 2019). Such English-centric trend of commonsense reasoning studies not only limits our research scope, but also tends to exacerbate English-specific bias that might prevent future methods from generalizing beyond English (Ponti et al., 2020).

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 1274–1287
August 1–6, 2021.

It is of pressing urgency for the community to develop NLU systems that can serve *all* languages in the world to bridge the gap between different cultures and eliminate language barriers (Hu et al., 2020), and *multilingual language models* (ML-LMs), such as XLM-R (Conneau et al., 2020), are among the most promising tools to achieve this ambitious goal. Although ML-LMs have been evaluated in a few NLU tasks, e.g., XNLI (Conneau et al., 2018) and XTEMRE (Hu et al., 2020), it is still relatively unclear how ML-LMs perform in commonsense reasoning tasks, due to the lack of 1) dedicated methods for probing common sense in ML-LMs and 2) multilingual benchmark datasets for commonsense reasoning.

To analyze how much common sense ML-LMs already have *without any tuning*, we propose MICKEYPROBE, a zero-shot probing task. It tasks a ML-LM to rank a set of *contrastive* assertions (i.e., declarative sentences) in the same language by their *commonsense plausibility*, for which we use *pseudo-likelihood* (PLL) (Salazar et al., 2020) as a proxy. Unlike the LAMA probe, it can study *multi-token concepts* which are ubiquitous in some non-English languages. In addition, it fairly compares performance across different languages via a *language-invariant* evaluation protocol. Alongside the probing task, we also create `MickeyCorpus`, a large-scale multilingual dataset, consisting of 561k sentences in 11 different languages. Our experiments reveal that there are always large discrepancies across different languages in the tested ML-LMs, and different ML-LMs show very different language preferences.

Beyond supervision-free analysis of ML-LMs, we also study their performance in commonsense reasoning tasks, such as CSQA and CODAH, within a *cross-lingual transfer* setting (i.e., trained on English data and tested on other languages). We find that existing ML-LMs tend to have much lower accuracy in commonsense reasoning beyond English. We conjecture a major common weakness of existing ML-LMs is that their pretraining stages do not have a proper *sentence-level* objective. Therefore, we propose *multilingual contrastive pre-training* (MCP), which tasks a ML-LM to select the correct assertion out of a set of N contrastive assertions in N *different* languages. We re-format `MickeyCorpus` by sampling across languages and thus form a dedicated pre-training corpus for the MCP task. To fairly evaluate different ML-LMs and validate the effectiveness of MCP, we create X-CSQA and X-CODAH, two cross-lingual commonsense reasoning datasets by translating their English versions to 15 other languages[2], including low-resource ones such as Swahili (*sw*) and Urdu (*ur*). Experiments show that the proposed MCP objective indeed significantly improves the performance of state-of-the-art ML-LMs in cross-lingual commonsense reasoning. Our contributions are as follows:

- **Resources.** We collect a large multilingual parallel corpus, `MickeyCorpus`, consisting of 561k sentences in 11 languages, which can be used for *analyzing* and *improving* ML-LMs. We also create `X-CSQA` and `X-CODAH`, two cross-lingual CSR benchmarks in 16 languages, for question answering and scene completion, respectively.
- **Evaluation and analysis.** We analyze multiple popular ML-LMs with MICKEYPROBE, a *language-invariant*, zero-shot task for probing common sense in ML-LMs; We also evaluate them on X-CSQA and X-CODAH in a cross-lingual transfer setting.
- **Method to improve ML-LMs.** We propose *multilingual contrastive pretraining*, a simple and effective sentence-level pretext task for enhancing ML-LMs in cross-lingual commonsense reasoning, which significantly improves the state-of-the-art ML-LMs in cross-lingual commonsense reasoning.

2 Background and Related Work

In this section, we introduce important concepts, background knowledge, and related work before we present our work in following sections.

2.1 Multilingual Language Models

A multilingual language model (ML-LM) aims to produce text representations for multiple languages in a unified embedding space. One of the unique advantages of ML-LMs is their potential ability to perform **zero-shot cross-lingual transfer** — a model trained (or fine-tuned) on data in one language (usually English) can be directly used in other languages as well without further fine-tuning. Improving ML-LMs is thus believed as one of the most promising approach towards multilingual NLU at scale. mBERT (Devlin

[2]The **16 languages** for X-CSQA and X-CODAH: {en, zh, de, es, fr, it, jap, nl, pl, pt, ru, ar, vi, hi, *sw*, *ur*}.

et al., 2019) is simply the BERT model (Devlin et al., 2019) trained on multilingual corpora without specific designs about multilinguality. The distil-mBERT (d-mBERT) (Sanh et al., 2019) is a smaller mBERT trained by knowledge distillation. Conneau and Lample (2019) proposed XLM(-100), which is pretrained with both masked language modeling (MLM) and translation language modeling (TLM). Conneau et al. (2020) further proposed XLM-R, which improves the XLM with a better sub-token vocabulary and high-quality multilingual corpora (CC100). We leave the analysis of recent seq2seq ML-LMs, such as mBART (Liu et al., 2020) and mT5 (Xue et al., 2021), as future work, because their architectures are significantly different from the other ML-LMs.

Note that the above ML-LMs are pretrained only with **token-level** training objectives such as MLM (i.e., recovering masked tokens in monolingual text) and TLM (i.e., recovering masked tokens in a pair of parallel sentences in two different languages). However, most NLU tasks, including commonsense reasoning, highly rely on **sentence-level** representations. We argue that a well-designed sentence-level pre-training objective should improve ML-LMs for NLU tasks. This intuition motivates us to propose a sentence-level pre-training objective — MCP (Section 5).

2.2 Cross-lingual Language Understanding

There are a few recent multilingual benchmarks for NLU tasks, e.g., XTREME(Hu et al., 2020), TyDi QA(Clark et al., 2020), and XGLUE(Liang et al., 2020). XTREME and XGLUE are unified large-scale multilingual multitask benchmarks, while Ty-Di QA focuses on the QA. These existing cross-lingual benchmarks have not covered *commonsense reasoning tasks*, such as CSQA (Talmor et al., 2019), SWAG (Zellers et al., 2018), and CODAH (Chen et al., 2019).

CSQA is a question answering task and the other two are scene completion tasks, while all have a multiple-choice selection objective, as shown in Figure 1. These benchmarks are widely used to evaluate LMs for commonsense reasoning. Unfortunately, they are limited to English, not applicable to evaluate models of multilingual commonsense knowledge, which motivates us to create X-CSQA and X-CODAH. The goal of the recent XCOPA (Ponti et al., 2020) dataset shares a similar goal, but it only focused on event-based causal reasoning in the scope of humans' social behavior, which is thus arguably more culturally biased. In contrast, the X-CSQA and X-CODAH are mainly for evaluating general world knowledge and cover more fine-grained types of reasoning (e.g., quantitative, negation), and thus engage a more language-agnostic, comprehensive understanding of ML-LMs about common sense.

2.3 The LAMA Probe and Its Limitations

The LAMA Probe (Petroni et al., 2019) is the seminal work on probing for common sense in (English) language models. It has a straightforward intuition: if a pretrained language model contains more commonsense knowledge, then it should be better at recalling a masked token in a commonsense assertion (e.g.,"*birds have [mask]*"). Specifically, given a LAMA-probe sentence $\boldsymbol{s}$ and its masked token w_t, a LM under testing uses all past and future tokens — $\boldsymbol{s}_{\setminus t} := \left(w_1, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{|\boldsymbol{s}|}\right)$. as the input to rank all tokens in the vocabulary with the probability $P\left(w_t \mid \boldsymbol{s}_{\setminus t}\right)$ via *zero-shot* inference. One can evaluate the performance of recalling common sense by measuring the position of a correct token "wing" in the ranked list. That is, the LAMA probe method uses **token-level probability** as a proxy to probe for common sense in LMs via ranking all tokens in their vocabularies.

This intuitive method, however, has several inherent limitations. First, in many other languages, *multi-token concepts* are ubiquitous, for example, "图书馆" ("library" in Simplified Chinese). Jiang et al. (2020) present several methods to decode multi-token entities so that they can adapt the LAMA probe to probe a LM for *language-specific* analysis. It is however infeasible to use token-level probing tasks if we want to analyze ML-LMs *across languages*. In addition, the evaluation metric of the LAMA probe could be unfair, because there can be many correct words for a masked position (e.g., "*birds have legs/eyes*"). The ranking metrics of the LAMA probe, however, tend to ignore these facts, resulting in a less trustworthy analysis. The vocabulary-specific ranking is unfair when comparing across different languages, so they can have very different label space. These limitations of the LAMA Probe prevent us from analyzing common sense in ML-LM across topologically diverse languages.

3 The Mickey Probe

The challenges of using the LAMA Probe for probing common sense in ML-LMs motivate us to propose a more suitable method for analyzing ML-LMs, one that can fairly compare across a diverse set of languages. We present MICKEYPROBE, a *Mul*t*i*lingual task for probing *c*ommonsense *k*nowledg*e* and analysis. We design a language-agnostic probing task with a sentence-selection objective for analyzing common sense of a ML-LM: given a set of assertions (i.e., declarative sentences) that have similar words and syntactic features, select the one with highest commonsense plausibility. We present the task formulation in this section and then introduce how we collect the dedicated dataset in Section 4.

Notations. We define a Mickey probe M as a set of K assertions in the same language, where one and only one of them (say, M_i) is the truth assertion with better commonsense plausibility than the other $K-1$ ones. Each Mickey probe M has multiple semantically equivalent versions in different languages. Let us denote a *language* by $l \in \mathcal{L}$ where $\mathcal{L} = \{en, fr, ru, zh, \dots\}$ and $|\mathcal{L}|$ is the number of languages of interest. Then, M^l is the probe M in the language l. For example, M^{en} and M^{fr} denote the probes with the same meaning but in English (en) and French (fr) respectively. We use $\mathcal{M}$ to denote a multilingual parallel dataset for MICKEYPROBE, which consists of $T \times |\mathcal{L}| \times K$ assertions. T is the number of MICKEYPROBE items and each item has K assertions and $|\mathcal{L}|$ language. Finally, we can formally describe a multilingual parallel dataset $\mathcal{M}$ for MICKEYPROBE:

$$\forall M \in \mathcal{M},\ \forall (l_x, l_y) \in \mathcal{L}^2,\ \forall i \in \mathbb{N}_{\le K},\ M_i^{l_x} \bowtie M_i^{l_y}\ . \quad (1)$$

We use the notation $\bowtie$ to indicate two assertions in different languages (e.g., l_x and l_y) are semantically equivalent to each other. We leave the details of creating such an $\mathcal{M}$ in Section 4.

Commonsense Probing Task. Given a Micky Probe M in the dataset $\mathcal{M}$, and suppose the index of the truth assertion to be t, a perfect multilingual language model would produce sentence probabilities such that it always gives the truth assertion M_t^l the highest probability among other candidates for every language.

$$\forall l \in \mathcal{L}, \forall i \in \mathbb{N}_{\le K},\ P(M_i^l) \le P(M_t^l). \quad (2)$$

Figure 2: A Mickey Probe example M has a set of probes in different languages (e.g., $M^{\text{en/zh}}$), and each of them is a set of 5 assertions. We rank assertions in the same language by their PLLs to probe common sense in ML-LMs across different languages.

It is still an open problem to properly compute sentence probabilities from masked language models, the recently proposed *pseudo-log-likelihood scoring* (PLLs) (Salazar et al., 2020) has shown promising results in many downstream NLP applications that need sentence re-ranking (e.g., speech recognition, and translation), suggesting it is a promising proxy of sentence probability. Given a sentence $\boldsymbol{s}$, its PLL is defined as:

$$\log P(\boldsymbol{s}) = \text{PLL}(\boldsymbol{s}) := \sum_{i=1}^{|\boldsymbol{s}|} \log P\left(w_i \mid \boldsymbol{s}_{\setminus i}\right) \quad (3)$$

That is, we individually mask each token w_i at a time and use the remaining context $\boldsymbol{s}_{\setminus i}$ to get the probability of a word $\boldsymbol{w_i}$ in the sentence $\boldsymbol{s}$. Finally, we aggregate them to approximate $P(\boldsymbol{s})$.

Evaluation Metric. The evaluation metric for MICKEYPROBE over a multilingual parallel dataset $\mathcal{M}$ in a specific language l is defined as the overall hit@k accuracy of the selection results $\text{hit@}k\,(l) = \sum_{M \in \mathcal{M}} \mathbb{1}\{\text{truth-rank}(M^l) \le k\}\ /\ |\mathcal{M}|$ where $\text{truth-rank}(M^l)$ means the the position of the truth assertion M_t^l in M^l sorted by their probabilities defined in Eq. (3). The hit@1 is just equivalent to the conventional *accuracy*.

Advantages of MICKEYPROBE. There are two key advantages of the MICKEYPROBE for evaluating ML-LMs: (1) The *sentence-level probability* can be more generally applied in languages besides English, comparing with the LAMA probe which only studies single-token English words.

Models \ $\mathcal{L}$	en	de	it	es	fr	nl	ru	bg	vi	zh	hi	avg
BT-Cosine	1.0	0.937	0.936	0.935	0.934	0.933	0.901	0.901	0.882	0.879	0.869	0.919
CC-size (GB)	300.8	66.6	30.2	53.3	56.8	29.3	278.0	57.5	137.3	46.9	20.2	97.9
Shortest	23.17	27.21	29.93	31.00	35.84	31.68	18.55	22.01	15.46	25.07	20.66	25.51
d-mBERT	62.95	34.56	25.26	34.85	50.46	32.39	21.49	29.14	19.77	32.57	25.88	33.57
mBERT	63.56	35.58	29.13	44.70	42.58	35.15	28.30	36.03	24.04	28.15	27.85	35.92
XLM-100	60.57	36.33	26.49	43.39	32.53	36.24	32.90	39.71	25.79	33.01	31.49	36.22
XLM-R$_B$	89.69	58.94	53.45	60.88	49.12	59.99	45.74	45.26	41.65	51.02	40.73	54.22
XLM-R$_L$	90.03	61.98	53.42	63.68	59.47	63.12	50.03	47.01	45.30	55.93	43.98	57.63

Table 1: The hit@1 accuracy (%) of the five ML-LMs for the MICKEYPROBE task.

(2) The task formulation creates a relatively closed-ended setting, such that we can use a *language-independent evaluation metric* to fairly compare across various languages within a ML-LM and compare across various ML-LMs for a particular language. In addition, we can see LAMA Probe as a *monolingual, word-level* version of the more general MICKEYPROBE: the LAMA Probe is when $\mathcal{L} = \{en\}$, and $\{M^{\text{en}}\} = M \in \mathcal{M}$ is a *huge* number of K assertions (i.e., the vocabulary size) — a *fixed* `[mask]` is replaced by all tokens in the vocabulary.

4 The Mickey Corpus and Evaluation

We present a procedure for automatically creating a multilingual parallel dataset $\mathcal{M}$ for the probing task MICKEYPROBE. Our collected corpus, named `MickeyCorpus` , has 561k sentences in 11 languages (T =10.2k, K=5, $|\mathcal{L}|$=11).

4.1 Creating English Probes

For the correct commonsense assertions in English, we have an existing resource, the OMCS corpus (Singh et al., 2002) which contains human-written sentences in English that describe commonsense facts. Each assertion can be used as a M_t^{en} and we perform perturbations on it to create the other $K-1$ distractor assertions (i.e., false candidates), yielding an M^{en} example.

Inspired by BERT-attack method (Li et al., 2020), we use a simple method to generate false assertions that are semantically related and syntactically similar to the truth assertions. Given a correct assertion, we first randomly sample a few ($1 \sim 3$) words with a part-of-speech tag as noun, verb, or adjective, and replace them with [mask]. Then, we use a beam-search style method to decode the [mask] tokens one by one from left to right. To ensure that the distractors are less plausible, we limit the decoding steps to only sample tokens that ranks between 200th∼300th. We repeat the above procedure multiple times with different sets of `[mask]` tokens. Then, we use Stanza (Qi et al., 2020) to remove distractors that have sequences of POS tags or morphological features different from the truth assertions. Finally, we sample $K-1$ of them as the distractors.

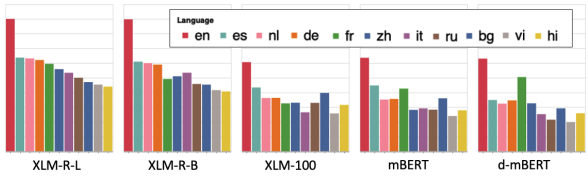

Figure 3: The MICKEYPROBE results in hit@1-acc. A larger version of this figure is in Appendix (Fig. 6).

4.2 Scaling to Ten Other Languages.

We use *bidirectional translation* with the MarianMT models (Junczys-Dowmunt et al., 2018) pretrained on the OPUS corpora (Tiedemann, 2016). We translate all English probes to the 25 languages that has models in both directions and then translate them back to English. As the outputs from these models might contain noise and errors, we compute the semantic similarities (i.e., cosine similarity) between the original M^{en} and the back-translated $M^{\text{x-en}}$ via the SentenceBERT (Reimers and Gurevych, 2019) model.

To ensure the quality and fair comparisons, we set a similarity threshold as 0.75 and keep the intersections of probes in all languages. Considering some languages tend to have translations of lower quality, we finally choose the best 10 languages to build the Mickey Probe dataset for our analysis, yielding 10k examples in each language and 10.2k*5*11 $\approx$ 561k sentences in total. The language set $\mathcal{L}$ =

$\{en, de, fr, ru, es, hi, vi, bg, zh, nl, it\}$.

Note that our purpose of checking the back-translation quality here is mainly to only keep the high-quality translations for all language pairs that we considered. Conventional metrics, e.g., BLUE score (Papineni et al., 2002), which focus on the *exact* word match, are thus less suitable: given the original sentence "I have a book", the translation results "I have a novel" and "I have a tool" will be seen as equally wrong. Inspired by BERTScore (Zhang et al., 2020), the BT-cosine is based on SentenceBERT, which efficiently gives a higher score for the former and a lower score for the latter, due to the semantic relatedness between "novel" and "book." We observed that most of our back-translations are in similar situations, and thus decide to use BT-cosine instead of others.

4.3 Analyzing ML-LMs with Mickey Probes

We now use the `MickeyCorpus` to evaluate the 5 pre-trained ML-LMs introduced in Section 2.1: d-mBERT (Sanh et al., 2019), mBERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019), XLM-R_{Base}, and XLM-R_{Large} (Conneau et al., 2020). All these ML-LMs pretraining objectives contain masked-word-prediction tasks, so we can easily use PPLs (Eq. 3) to probe them *a zero-shot, supervision-free manner* with hit@1 accuracy. (The hit@2 results are shown in Appendix.) We present a histogram in Figure 3 and show the concrete results in Table 1. We find that there are always large discrepancies across different languages in all tested ML-LMs, which motivates us to analyze the following questions.

Q1: Do different ML-LMs have similar language preferences? No. We arrange the languages in all ML-LMs with the same order for Figure 3 — the *monotonically* descending order of XLM-R_L. Interestingly, we find that different ML-LMs are good for different languages, resulting in a very diverse set of trends. For example, XLM-R_B, has a higher performance in *it* than *zh* and *fr*, unlike XLM-R$-L$ which are pre-trained on the same corpora with the same objectives. mBERT and d-mBERT has stronger performance in *fr* than *nl* and *de*, unlike XLM and XLM-R.

Q2: Does length influence PLL ranking? Not much. The PLL computation indeed tends to prefer shorter sequences (see Eq. 3), so one may wonder if the length of assertions would influence the probing results. The "Shortest" row in Table 1 presents the results when we always select the shortest assertion within a probe, instead of PLL ranking. The gaps between these scores and XLM-R-L's suggest that the probing task indeed uses PLL as a valid proxy for evaluating common sense based on sentence-level semantics.

Q3: Is the translation quality a key factor? We show "BT-Cosine", the mean of the cosine scores between the original English sentences and the back-translated ones, and sort the table by these numbers. The first 5 languages, {de, it, es, fr, nl} have the largest BT-Cosine, i.e., the best translation quality, and they indeed have better performances in general for XLM-R models. However, although *zh* has a worse BT-score than *vi*, all ML-LMs perform better in *zh* than *vi*. Thus, we believe the translation quality of `MickeyCorpus` will not be a factor to influence our understanding of ML-LMs. Consequently, this suggests that further study must depend on pre-training corpora of each ML-LM in different languages.

Q4: Does the size of pre-training corpora matter? We list the size of the monolingual corpus in each language for CC-100 that XLM-R are pre-trained on (i.e., the CC-size row). Although *ru* has a much larger corpus than *de*, *it*, etc., the XLM-R performance in *ru* is much worse. In addition, *fr* and *nl* have almost the same translation quality while *fr*'s CC-size is twice the size of *nl*, but the performance in *fr* is still much worse than *nl*. We conjecture this would be either due to the design of sub-token vocabulary or the text quality (instead of the size) of the CC-100 corpora.

Further implications. The benchmark results of five popular ML-LMs on the MICKEYPROBE task over the `MickeyCorpus` offer the initial and valuable understanding with a closer look at the commonsense knowledge of ML-LMs by probing them in a unified evaluation protocol. One can either compare a ML-LM across different languages or compare a certain language across ML-LMs in Table 1. These comparable results support further analysis that can benefit the development of ML-LMs in the future. After all, even the best ML-LM XLM-R_L also degrades much in other languages, and also perform slightly worse than RoBERTa$_L$ in *en* (93.4%). We argue (culture-invariant) common sense knowledge should be seen as an important way to connect multiple languages and thus better align them in a shared embedding space induced by a ML-LM.

5 Multilingual Contrastive Pre-Training

In this section, we reformat the MICKEYPROBE so that we can reuse the `MickeyCorpus` for improving the pre-trained ML-LMs for commonsense reasoning beyond English. We propose a *multilingual contrastive pre-training* (MCP) task that focuses on enhancing the sentence-level representation of ML-LMs. MCP improves a ML-LM in a *multilingual*, *contrastive* environment, where the model learns to select the assertion with the best commonsense plausibility from a set of contrastive sentences *in different languages*. Each MCP example is a set of *multilingual* assertions while each Mickey probe is a *monolingual* set.

MCP Dataset Creation from $\mathcal{M}$. We create pretraining examples for the MCP task by converting MICKEYPROBE examples, as shown in the steps illustrated in Algorithm 1. Simply put, we reformat a K-way Mickey Probe M ($K \times |\mathcal{L}|$ assertions) to a MCP example by sampling a set of V candidate assertions in V different languages. We convert all examples in the `MickeyCorpus` $\mathcal{M}$ to build a new *cross-lingual sentence-selection* dataset $\mathcal{C}$ for learning the MCP task.

MCP Learning. Given a MCP example $C \in \mathcal{C}$, we append one dense linear layer f on top of a ML-LM with parameters denoted as $\Theta_{\text{ML-LM}}$ for learning to predict the *commonsense plausibility score* of each assertion $C_i \in C$ as follows:

$$\mathbf{h_i} = \text{ML-LM}(C_i).\texttt{[CLS]} \quad (4)$$

$$o_i = f(\mathbf{h_i}; \Theta_f) \quad (5)$$

$$z_i = \frac{e^{o_i}}{\sum_{j=1}^{V=|C|} e^{o_j}} \quad (6)$$

$$\rho = \sum_{i=1}^{V} -\mathbb{1}_i \log(z_i) \quad (7)$$

We first get the logit o_i of each assertion by projecting its `[CLS]` embeddings $\mathbf{h_i}$ to a logit o_i via a dense layer f with parameters Θ_f; Then, we use SoftMax to normalize the logits as plausibility scores z_i; Finally, we compute the cross-entropy loss ρ where $\mathbb{1}_i$=1 if C_i is a correct assertion and 0 otherwise. We fine-tune $\{\Theta_{\text{ML-LM}}, \Theta_f\}$ to minimize the overall loss over the MCP dataset $\mathcal{C}$.

6 Evaluation for Cross-lingual CSR

In this section, we introduce the datasets, experimental setup, results, and our analysis.

Algorithm 1: Convert a Mickey Probe M to an example for the MCP task.

In: $M \in \mathcal{M}$ /* is a probe that has $|\mathcal{L}|$ sub-sets; each sub-set M^{l_x} is a set of K assertions in the same language $l_x \in \mathcal{L}$. $M_t^{l_x}$ is always the truth. */

Out: C /* A set of V assertions in different languages. */

Remarks: $\Gamma_n(X)$ is a function to randomly sample n unique elements from a set X.

$l_a \leftarrow \Gamma_1(\mathcal{L})$ /* Pick an anchor language. */

$C \leftarrow \{M_t^{l_a}\}$ /* Initiate w/ the truth assertion. */

/* Iterate each sampled distractor language l_i. */

foreach $l_i \in \Gamma_{V-1}(\mathcal{L} - l_a)$ **do**

  /* Sample an index of distractor assertion. */

  $j \leftarrow \Gamma_1(\mathbb{N}_{\leq K} - \{t\})$

  /* Add a distractor assertion as a candidate. */

  C.add($M_j^{l_i}$)

6.1 X-CSQA & X-CODAH: Two New Benchmarks for Evaluating XCSR

To evaluate ML-LMs for commonsense reasoning in a cross-lingual zero-shot transfer setting, we create two benchmark datasets, namely X-CSQA and X-CODAH. Table 3 shows the statistics of the two datasets. Specifically, we use online commercial services such as *DeepL Pro Translate* to collect high-quality translations of the examples in CSQA and CODAH for **15 languages** other than English. The size of CODAH is small (only 2.7k), so we use 7k SWAG validation examples as additional training data which share the same formulation. We discuss the reduction of *cultural differences* and quality control of automatic translations as well as other details in *Ethical Considerations* (the paragraph for cultural bias reduction) and *Appendix* (A). As our goal is to evaluate different ML-LMs (instead of different languages) in a unified evaluation protocol for cross-lingual commonsense reasoning, we argue that such automatically translated examples, although might contain noise, can serve as a starting benchmark for us to obtain meaningful analysis before more human-translated datasets will be available in the future.

6.2 Setup

We focus on 4 popular ML-LMs that we introduced in Section 2.1: mBERT, XLM-100, XLM-R_B and XLM-R_L as well as our proposed MCP method. For both tasks, we concatenate each prompt (the question or first sentence) and each

	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

Table 2: Benchmark results for different ML-LMs and MCP-enhanced models for X-CSQA and X-CODAH in a zero-shot cross-lingual setting. Δ is the improvement of MCP. {*pl,ar,ja,pt,sw,ur*} are unseen in MCP.

Stat. ↓ Dataset →	X-CSQA	X-CODAH
Task Format	QA	SceneComp.
# Languages	15	15
# 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

Table 3: Statistics of the two X-CSR datasets.

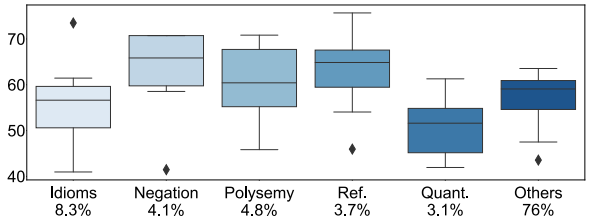

Figure 4: Categorized accuracy in for MCP(XLM-R$_L$) on X-CODAH. Each box is for 15 languages.

of its options individually in the form of "[CLS] prompt [SEP] option$_i$ [SEP]". Then, we fine-tune ML-LMs over the English training dataset and test them on other languages.

Why zero-shot cross-lingual transfer? It is almost impossible to collect data in *all* languages that an NLU system might be used for. Therefore, prior works mainly focus on zero-shot cross-lingual transfer (Conneau et al., 2018), which is more meaningful and can offer *lower-bound* performance analysis. It is also an ideal setting for studying CSR because most commonsense facts are *language-invariant*. Thus, an English-finetuned ML-LM for CSR should be able to transfer its ability to a wide range of other languages as well. Furthermore, our goal of this paper is to evaluate and improve ML-LMs, so translating back to English and then use an English-only LM is also not helpful towards to this end.

6.3 Experiments for Cross-lingual CSR

In Table 2, we present the empirical results over X-CODAH and X-CSQA for the ML-LMs as well as two models enhanced by our proposed MCP method. On both tasks, the XLM-R$_L$ performs the best with a large margin. Enhanced by the MCP method, both XLM-R$_B$ and XLM-R$_L$ see significant improvement (e.g., 2.7% absolute improvement for XLM-R$_L$ on X-CSQA-avg).

Can MCP's improvement generalize to unseen, low-resource languages? Note that MCP dataset only involves 9 languages here, and there are 6 languages that are totally *unseen* in the MCP training (i.e., {*pl, ar, ja, pt, sw, ur*}). The largest performance gain is in *ru* on X-CSQA and *vi* on X-CODAH. Surprisingly, we find the improvements on them are also large for XLM-R$_L$ (e.g., 48.4→ 52.3 for *ar*). In addition, for the two *low-resource* languages *sw* and *ur*, MCP also brings 2 $\sim$ 3 percentage points of improvement for XLM-R$_L$. It is, however, not always the case for *XLM-R$_B$*, which we conjecture tends to be more likely to overfit.

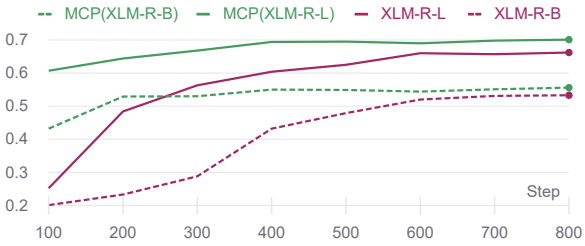

Figure 5: Dev acc v.s. learning steps on X-CSQA.

Although ML-LMs enjoy the merits of zero-shot cross-lingual transfer, their performances are usually *worse* than the English-only RoBERTa$_L$ on the en-test (70.4% vs 66.7% for X-CSQA). Although MCP can mitigate the gap (70.4% vs 69.5%) for X-CSQA, there is still a large gap (81.6% vs 69.9%) for X-CODAH. We use Fig. 4 to analyze how different categories of commonsense reasoning in CODAH (Chen et al., 2019) are diverse in different languages. We find that *others*, *reference*, and *negation* have relatively smaller variances across different languages, as they are more language-invariant. However, a few *polysemous*, *idioms* examples can be English-specific which may not generalize to other languages. More detailed analysis is in Appendix.

From the curve of dev accuracy in Figure 5, we see that MCP-enhanced XLM-R models are much more *sample efficient* and converge much faster than vanilla versions. This suggests that the MCP, if used on a larger corpus with broader topics, can potentially produce a better ML-LM with more general usage, especially when only limited labelled is available. Our results on XNLI-10% (using 10% of the training data) (Conneau et al., 2018) show that MCP-enhanced XLM-R$_L$ has 1.2 percent accuracy improvement on the average of 15 languages. As our focus in this paper is commonsense reasoning, we leave the study on other cross-lingual NLU tasks as future work. Importantly, our experiments imply that a proper (continual) pre-training task that has a (contrastive) sentence-level objective could improve both the final performance as well as learning efficiency.

7 Conclusion

We evaluate and improve popular multilingual language models (ML-LMs) for advancing commonsense reasoning beyond English. We propose the MICKEYPROBE, a *language-agnostic* probing task for analyzing common sense of ML-LMs in a zero-shot manner. With our proposed new benchmark datasets via automatic translation, X-CSQA and X-CODAH, we evaluate ML-LMs in a cross-lingual transfer setting for commonsense reasoning. We also improve the state-of-the-art ML-LM with a simple yet effective method — multilingual contrastive pre-training, which uses a sentence-level objective to enhance sentence representations, yielding a significant performance gain. All above work is based on `MickeyCorpus`, which can be used as both a probing dataset and a pre-training corpus for analyzing and improving ML-LMs. We hope our resources and pre-training method for ML-LMs can help the community advance commonsense reasoning beyond English.

Acknowledgements

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Contract No. 2019-19051600007, the DARPA MCS program under Contract No. N660011924033 with the United States Office Of Naval Research, the Defense Advanced Research Projects Agency with award W911NF-19-20271, and NSF SMA 18-29268. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. We would like to thank all the collaborators in USC INK research lab and the reviewers for their constructive feedback on the work.

* Ethical Considerations

Resource Copyright This work presents three new resources: `MickeyCorpus`, X-CODAH, and X-CSQA, which are multilingual extension of the OMCS (Singh et al., 2002) [3], CSQA (Talmor et al., 2019)[4], and CODAH (Chen et al., 2019)[5] respectively. All these three original sources of the data are publicly available for free, and we do not add any additional requirement for accessing our resources. We will highlight the original sources of our data and ask users to cite the original papers when they use our extended versions for research.

[3] https://github.com/commonsense/conceptnet5/wiki/Downloads

[4] https://www.tau-nlp.org/commonsenseqa

[5] https://github.com/Websail-NU/CODAH

Cultural Bias Reduction Like most most multilingual parallel resources, especially in general NLU domain, there exists potential data bias due to the barrier of languages as well as *cultural differences* (Acharya et al., 2020; Lin et al., 2018), which could induce the labeling differences on the same situation. For example, a question like "what do people usually drink in the morning? (coffee/tea/milk)" or "when does a wedding usually start? (morning/afternoon/evening)" might be answered very differently by people from different backgrounds and cultures, not to mention different languages. The prior English commonsense resources which our datasets are built on are already possess such inherent bias, even with in the English language. Therefore, before we translate CSQA and CODAH, we intentionally remove the examples that are either labeled as non-neutral by a pre-trained sentiment classifier, or contained any keywords that are relevant to social behavior (e.g., weddings). We manually inspect test examples in X-CSQA and X-CODAH in the English and Chinese versions and have a strong confidence there is few strongly controversial example. However, we admit that such reduction of cultural differences in common sense has not been systematically measured in this work for other languages.

Application Risks of Cross-lingual CSR.

The work also evaluates a few multilingual language models (ML-LMs) for cross-lingual commonsense reasoning (XCSR), and introduced a new model which outperforms them. This raises the question of whether harm might arise from applications of XCSR—or more generally, since XCSR is intended as a step toward making English-only CSR more applicable in other languages, whether harm might arise more generally from existing ML-LMs. Among the risks that need to be considered in any deployment of NLP technology are that responses may be wrong or biased, in ways that would lead to improperly justified decisions. Although in our view the current technology is still relatively immature, and unlikely to be fielded in applications that would cause harm of this sort, it is desirable that ML-LMs provide audit trails, and recourse so that their predictions can be explained to and critiqued by affected parties.

References

A. Acharya, Kartik Talamadupula, and Mark A. Finlayson. 2020. An atlas of cultural commonsense for machine reasoning. *ArXiv*, abs/2009.05664.

Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. 2019. CODAH: An adversarially-authored question answering dataset for common sense. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 63–69, Minneapolis, USA. Association for Computational Linguistics.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9):92–103.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. 2020. Scalable multi-hop relational reasoning for knowledge-aware

question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1295–1309, Online. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization. Technical report.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. 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)*, pages 2829–2839, Hong Kong, China. Association for Computational Linguistics.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6862–6868, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Frank F. Xu, Kenny Zhu, and Seungwon Hwang. 2018. Mining cross-cultural differences and similarities in social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 709–719, Melbourne, Australia. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 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)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Push Singh, Thomas Lin, Erik T Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. Open mind common sense: Knowledge acquisition from the general public. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 1223–1237. Springer.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Jörg Tiedemann. 2016. OPUS – parallel corpora for everyone. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*, Riga, Latvia. Baltic Journal of Modern Computing.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Appendix

A Details for Dataset Construction

Before we start the translation procedure, we first re-split the datasets of CSQA and CODAH such that the test set examples in the English language do not contain controversial examples or culture-related examples that would potentially cause cultural bias in our dataset. Please refer to the section of Ethical Considerations (following the Conclusion) in the main paper for more details. Then, we use the DeepL Pro translation service to translate the 10 languages: {de, fr, es, pt, it, nl, pl, ru, jap, zh} and use Google Translation API to translate the others {ar, sw, ur, vi, hi}.

We agree that ideally we should use human experts to translate the examples in CSQA and CODAH, but the cost or building a large-scale multilingual dataset with the same scale of our datasets is extremely high – around 10k USD. As a matter of fact, most of the examples in CSQA and CODAH are very easy and short sentences, and most of them can be well translated by modern commercial translation APIs, because they usually have a hybrid system. Moreover, we choose the DeepL online service because it has been reported by many individual media as the best choice. To ensure the quality of the translation, we perform the translation for both directions and then use the same quality control method as we discussed in Section 4 for removing the examples that have lower cosine similarity between original English version and back-translated examples. During the process, we manually went through the Chinese versions to find a suitable threshold for taking the intersection — 0.85, which results in a comparable BT-cosine mean to the XNLI dataset [6].

Models	#lgs	tnz	L	H_m	H_{ff}	A	V	#para
mBERT	104	WP	12	768	3072	12	110k	172M
XLM-100	100	BPE	16	1280	5120	16	200k	570M
XLM-R_B	100	SPM	12	768	3072	12	250k	270M
XLM-R_L	100	SPM	24	1024	4096	16	250k	550M

Table 4: Model Architectures.

B Hyper-parameters

We summarize hyper-parameters that we used for training ML-LMs on X-CODAH and X-CSQA in Table 7. *Evaluation Steps* are equally 100 for all models and datasets. *Maximum Sequence Length* is 100 for X-CODAH and 64 for X-CSQA. The batch size here refers to "*train batch size per device* $\times$ *# GPUs* $\times$ *# gradient accumulation steps*". Note that the MCP methods use the exactly the same hyper-parameters which we have found optimal by tuning over the dev set. The learning rates we tried for all models are from the range {3e-5, 2e-5, 1e-5, 8e-6, 6e-6, 5e-6}. The warm up steps are selected from {50, 100, 200, 300, 500}.

[6] We sampled 1k examples in the test set and follow the same procedure for the intersection language set.

C Details of ML-LMs

Table 4 shows the model architectures and sizes that we used from (Conneau et al., 2020). We show the tokenization (**tnz**) used by each Transformer model, the number of layers L, the number of hidden states of the model H_m, the dimension of the feed-forward layer H_{ff}, the number of attention heads A, the size of the vocabulary V and the total number of parameters #params.

D Additional Experimental Results

D.1 Hit@1 Accuracy in Histogram

D.2 Hit@k Accuracy of Mickey Probes

Table 5 shows the Hit@2 Accuracy of the five ML-LMs for the *MickeyProbe*. Hit@2 Accuracy evaluates whether the models can rank the correct assertion within top 2. Unlike Hit@1 which only accepts best predictions, Hit@2 is more flexible. Thus, the performances in Hit@2 increase compared to the ones in Hit@1. We can see that the discrepancies across languages still exist.

D.3 Categorized X-CODAH Analysis

Please refer the CODAH (Chen et al., 2019) paper for the definition and concrete examples in each category. We show benchmark results of MCP(XLM-R_L) on X-CODAH within different carriages in Table 6. The **RB** stands for using the RoBERTa-Large model to fine-tune on the English X-CODAH dataset. We find that the largest gaps in En are in the Idioms and the Others. Interestingly, we find that the quantities category is where MCP performs better than the RoBERTa large.

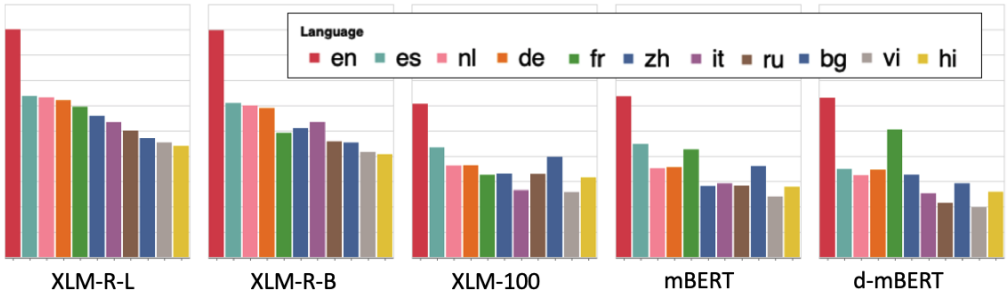

Figure 6: The MICKEYPROBE results in hit@1-acc. (An enlarged version of Figure 3.)

Models \ $\mathcal{L}$	**en**	**de**	**it**	**es**	**fr**	**nl**	**ru**	**bg**	**vi**	**zh**	**hi**	**avg**
Shortest	42.20	50.91	52.49	56.06	57.30	55.95	40.96	45.86	35.64	47.67	43.81	48.08
d-mBERT	87.06	61.48	47.70	62.30	76.17	59.03	45.71	55.47	42.53	60.24	52.56	59.11
mBERT	87.38	62.30	52.02	73.01	70.41	62.42	56.83	62.34	49.77	53.81	53.99	62.21
XLM-100	85.17	63.96	47.05	71.61	55.99	63.14	58.73	65.89	50.29	60.53	58.08	61.86
XLM-R$_B$	97.77	83.64	78.21	84.73	72.77	84.08	74.04	71.67	68.79	77.89	68.27	78.35
XLM-R$_L$	97.83	85.57	76.73	85.56	83.71	86.09	77.74	72.55	72.01	81.32	70.78	80.90

Table 5: The hit@2 accuracy of the five ML-LMs for the Mickey Probe task.

Category	**RB**	**en**	**de**	**it**	**es**	**fr**	**nl**	**ru**	**vi**	**zh**	**hi**	**pl**	**ar**	**ja**	**pt**	***sw***	***ur***	**avg**
Idioms	79.52	69.88	61.45	56.63	60.24	73.49	60.24	57.83	50.6	55.42	45.78	59.04	50.6	50.6	56.63	44.58	40.96	55.87
Neg.	75.61	75.61	65.85	65.85	70.73	70.73	58.54	70.73	65.85	70.73	63.41	65.85	60.98	58.54	70.73	41.46	58.54	64.63
Poly.	79.17	75.00	58.33	66.67	68.75	70.83	60.42	66.67	68.75	56.25	54.17	60.42	45.83	66.67	68.75	45.83	50	61.46
Ref.	86.49	78.38	62.16	67.57	67.57	64.86	64.86	67.57	62.16	54.05	67.57	72.97	75.68	45.95	54.05	62.16	56.76	64.02
Quant.	61.29	67.74	45.16	45.16	51.61	54.84	61.29	51.61	61.29	45.16	54.84	58.06	41.94	41.94	54.84	51.61	51.61	52.42
Others	82.89	68.95	61.05	62.37	59.74	59.08	60.66	57.37	63.03	63.55	53.29	57.89	54.08	55.13	60.79	43.55	47.5	58.00

Table 6: Benchmark results for MCP(XLM-R-L) on X-CODAH in different categories. RB = RoBERTa-Large.

Model	**lr**	**# epoch**	**# wus**	**bsz**
X-CODAH				
mBERT	3E-05	20	100	128
XLM-100	1E-05	20	100	64
XLM-R-B	1E-05	20	100	128
XLM-R-L	6E-06	10	100	64
MCP(XLM-R-B)	1E-05	20	100	128
MCP(XLM-R-L)	6E-06	10	100	64
X-CSQA				
mBERT	3E-05	30	100	64
XLM-100	1E-05	20	300	64
XLM-R-B	1E-05	30	100	144
XLM-R-L	6E-06	10	100	64
MCP(XLM-R-B)	1E-05	30	100	144
MCP(XLM-R-L)	6E-06	10	100	64

Table 7: The optimal hyper-parameters for fine-tuning. (**lr** represents 'learning rate'; training **# epoch** ; **# wus** = '# warm up steps'; **bsz** = 'batch size')