mCSQA: Multilingual Commonsense Reasoning Dataset with Unified Creation Strategy by Language Models and Humans

Yusuke Sakai, Hidetaka Kamigaito, Taro Watanabe
Nara Institute of Science and Technology
{sakai.yusuke.sr9, kamigaito.h, taro}@is.naist.jp

Abstract

It is very challenging to curate a dataset for language-specific knowledge and common sense in order to evaluate natural language understanding capabilities of language models. Due to the limitation in the availability of annotators, most current multilingual datasets are created through translation, which cannot evaluate such language-specific aspects. Therefore, we propose Multilingual CommonsenseQA (mCSQA) based on the construction process of CSQA but leveraging language models for a more efficient construction, e.g., by asking LM to generate questions/answers, refine answers and verify QAs followed by reduced human efforts for verification. Constructed dataset is a benchmark for cross-lingual language-transfer capabilities of multilingual LMs, and experimental results showed high language-transfer capabilities for questions that LMs could easily solve, but lower transfer capabilities for questions requiring deep knowledge or commonsense. This highlights the necessity of language-specific datasets for evaluation and training. Finally, our method demonstrated that multilingual LMs could create QA including language-specific knowledge, significantly reducing the dataset creation cost compared to manual creation. The datasets are available at https://huggingface.co/datasets/yusuke1997/mCSQA.

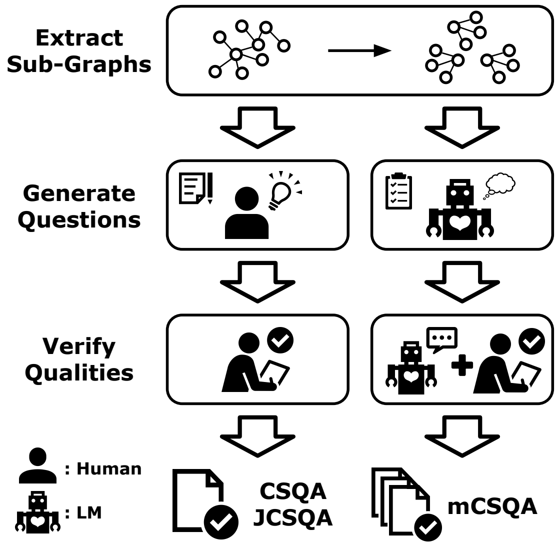

Figure 1: The comparison of the dataset creation process for mCSQA and (J) CSQA includes two key changes for efficient and low-cost creation of multilingual datasets. First, the question generation process shifts from human annotators to an LM. Second, an LM assists humans for the quality verification process.

Japanese Question (Translated to English)				
Q: お年寄りは？ (Who is the elderly person?)				
(a) わし (me)	(b) わたし (me)	(c) ぼく (me)	(d) おれ (me)	(e) うち (me)
English Question (Translated to Japanese)				
Q: How do we make a cake? (ケーキを作るにはどうする？)				
(a) roast (焼く) (burn)	(b) broil (焼く) (burn)	(c) grill (焼く) (burn)	(d) toast (焼く) (burn)	(e) bake (焼く) (burn)

Table 1: Examples require language-specific knowledge. They cannot be solved without such knowledge, as the translations consolidate the nuances into a single word.

1 Introduction

Can you choose the correct answer in Table 1? Each choice is semantically very close, making it difficult for non-native speakers to distinguish them. However, native speakers who have language-specific commonsense and knowledge can choose the most plausible choice considering subtle nuances. Despite the need to consider different backgrounds for each language, the datasets to evaluate the natural language understanding (NLU) capabilities of language models (LMs) are mostly for a few major languages such as English, and thus, many languages lack such datasets. When focusing on the cross-lingual capability of LMs, datasets created from scratch in multiple languages are lim-

Findings of the Association for Computational Linguistics ACL 2024, pages 14182–14214
August 11-16, 2024 ©2024 Association for Computational Linguistics

ited, and currently, evaluations mostly use datasets created through translation. However, as can be seen from the example in Table 1, datasets created through translation cannot accurately evaluate language-specific commonsense or knowledge. Therefore, it is necessary to create datasets for each language from scratch, but the manual creation of such datasets is limited by the availability of annotators and financial costs.

To tackle this problem, as shown in Figure 1, we propose a method to efficiently create multilingual NLU datasets from multilingual resources by replacing some of the manual annotation processes with generative multilingual LMs. In this study, we focus on CommonsenseQA (CSQA) (Talmor et al., 2019), a dataset for evaluating commonsense reasoning capabilities within NLU evaluations. CSQA is a major commonsense reasoning Question-Answering dataset manually created from the multilingual knowledge base ConceptNet (Speer et al., 2017). However, due to such limitations, CSQA has been created from scratch only in English and Japanese, JCommonsenseQA (JCSQA) (Kurihara et al., 2022). Therefore, we create a Multilingual CommonsenseQA (mCSQA) that extends CSQA to eight languages[1] using our proposed method.

Furthermore, we evaluated the cross-lingual language-transfer capabilities of multilingual LMs focusing on language-specific common sense and knowledge using mCSQA. The results showed high language-transfer capabilities for questions that LMs could easily solve, but lower transfer capabilities for questions requiring deep knowledge or commonsense. The total cost per question in mCSQA was reduced to one-hundredth of that for CSQA.

To summarize, our contributions are as follows:

- We propose an efficient and low-cost method for creating NLU datasets by generative multilingual LMs.
- We demonstrate the potential effectiveness of using multilingual LMs for creating datasets from multilingual resources.
- mCSQA makes it possible to analyze the cross-linguistic commonsense understanding capabilities and transfer performance from each language beyond English.
- The analysis revealed that, when focusing on language transfer capabilities using mCSQA, we identified cases where language-specific knowledge is required and cases where it is not, thereby confirming the need for non-translated language-specific datasets.

[1]English (en), Japanese (ja), Chinese (zh), German (de), Portuguese (pt), Dutch (nl), French (fr), Russian (ru)

2 Background and Related Work

Commonsense reasoning task This task evaluates how an LM can understand and infer object recognition, visual information, and cultural or societal common sense, which are not typically described in textual information. CSQA is a multiple-choice question task that asks for the most plausible choice as an answer with some variants: JCSQA is in Japanese, CommonsenseQA 2.0 (Talmor et al., 2021) is a more challenging dataset, ECQA (Aggarwal et al., 2021) requires explaining the process of deriving an answer, etc. There exist other types of commonsense tasks: COPA (Roemmele et al., 2011) and BalancedCOPA (Kavumba et al., 2019) ask about causal relationships between everyday events; SocialIQA (Sap et al., 2019b) asks about social common sense; PIQA (Bisk et al., 2020) evaluates procedural knowledge; HotpotQA (Yang et al., 2018) requires multi-hop inference; DROP (Dua et al., 2019) captures arithmetic operation capabilities; and tasks like understanding language information (Liu et al., 2022b; Kocijan et al., 2023; Sakaguchi et al., 2021; Wang et al., 2019), understanding causal relationships within documents (Mostafazadeh et al., 2020, 2016; Zhang et al., 2018; Huang et al., 2019; Ostermann et al., 2018; Smirnov, 2019), and CommonGen (Lin et al., 2020), which asks to generate common sentences from given keywords. The above datasets primarily focus on English, but there exist datasets in Japanese (Omura et al., 2020; Takahashi et al., 2019; Hayashibe, 2020), Chinese (Xu et al., 2021, 2020; Wang et al., 2022), Russian (Shavrina et al., 2020; Taktasheva et al., 2022), and Indonesian (Koto et al., 2022). For multilingual datasets, most are extended versions of existing ones through translation, such as X-COPA (Ponti et al., 2020) from COPA, X-CSQA (Lin et al., 2021) from CSQA, and X-CODAH (Lin et al., 2021) from CODAH (Chen et al., 2019). A few datasets, such as TyDiQA (Clark et al., 2020), are created for each language from scratch.

Multilingual datasets When focusing on the evaluation of multilingual performance of LMs,

Methods	Knowledge	Alignment	Costs
By translation	✗	✓	✓
Compilation of similar tasks	✓	✗	✓
From multilingual resources	✓	✓	✗
Ours	✓	✓	✓

Table 2: Categorize the multilingual datasets creation methods.

the evaluation datasets are almost exclusively created through three methods, as shown in Table 2: (1) Translation from existing datasets in a major language, e.g., English (Lin et al., 2021; Ponti et al., 2020; Conneau et al., 2018; Artetxe et al., 2020; Yang et al., 2019); (2) Compilation of similar tasks across multiple languages (Zhang et al., 2023c; Hu et al., 2023; Adelani et al., 2022; Roy et al., 2020; Malmasi and Dras, 2015); (3) Creation from multilingual resources following the same dataset creation process (Keung et al., 2020; Huang et al., 2020; Buchholz and Marsi, 2006; Clark et al., 2020; Schwenk and Li, 2018; Kabra et al., 2023). However, (1) translated datasets often do not account for language-specific culture, knowledge, common sense, or linguistic phenomena, leading to a bias towards the background of the source language (Hu et al., 2021; Lin et al., 2021; Acharya et al., 2020; Clark et al., 2020; Park et al., 2021; Kurihara et al., 2022). (2) Simply compiling datasets curated for each individual language could allow the evaluation of language-specific knowledge and common sense. However, it is difficult to align tasks across languages since most tasks differ in their creation methods data sources or philosophies. Thus, it just leads to evaluating the transfer capability among comparable tasks, and not evaluating the true transfer capabilities across languages. Therefore, (3) only the datasets created from multilingual resources can enable the evaluation of language transfer capability, considering the differences in language-specific knowledge and common sense. Nevertheless, the manual creation of such datasets is limited by the availability of annotators and financial costs.

Dataset creation with LMs The superior performance of generative language models allows to create datasets automatically. SWAG (Zellers et al., 2018) and HellaSwag (Zellers et al., 2019) have created answer choice options through the output of LMs. Such efforts have also been extended to use LMs for data augmentation (Staliūnaitė et al., 2021; Kumar et al., 2019, 2020; Lee et al., 2021). WANLI (Liu et al., 2022a), created from MNLI (Williams et al., 2018), employs GPT-3 (Brown et al., 2020) for adversarial data augmentation with manual checks to create challenging datasets. Some studies propose methods to manually check quality of LM generation results (Tekiroğlu et al., 2020; Yuan et al., 2021; Wiegreffe et al., 2022; Wang et al., 2021a; Li et al., 2023). Additionally, there are attempts to create datasets from scratch with emergent abilities of LMs, without using any examples (He et al., 2022; Wang et al., 2021b; Schick and Schütze, 2021; Meng et al., 2022; Ye et al., 2022). However, these studies have primarily focused on a single language, e.g., English. Recently, the outputs of language models themselves have been used to create datasets (Honovich et al., 2023; Shao et al., 2023; Sun et al., 2023; Peng et al., 2023) for instruction-tuning (Wei et al., 2022a). TarGEN (Gupta et al., 2023) employs a single language model and splits the data generation process into multiple steps, inputting the suitable prompt for each step to ensure data diversity and reliability. Putri et al. (2024) focus on middle-resource (Indonesian) and low-resource (Sundanese) languages, and investigate whether LLMs can create culturally aware commonsense questions by comparing translation datasets and those generated by LLMs from scratch.

3 Datasets Creation

Our mCSQA construction involves three main steps (see Figure 2): extraction of sub-graphs from ConceptNet, creation of question and choice pairs with LMs, and verification of their quality by both LMs and humans. We basically follow the creation processes of CSQA and JCSQA, but modified to allow for unified processing to support multiple languages.

3.1 Extract Sub-Graphs from ConceptNet

ConceptNet is a graph knowledge base defined as a tuple, $\mathcal{G} = (\mathcal{C}, \mathcal{R}, \mathcal{T})$, where $\mathcal{C}$ denotes a set of concept entities, $\mathcal{R}$ denotes a set of relations and $\mathcal{T}$ denotes a set of triples. Each triple is represented as $(s, r, t) \in \mathcal{T}$, where s and $t \in \mathcal{C}$ are the source and target concept entities, respectively, and $r \in \mathcal{R}$ is the relation, and carry commonsense knowledge such as "(student, CapableOf, forget to do homework)".

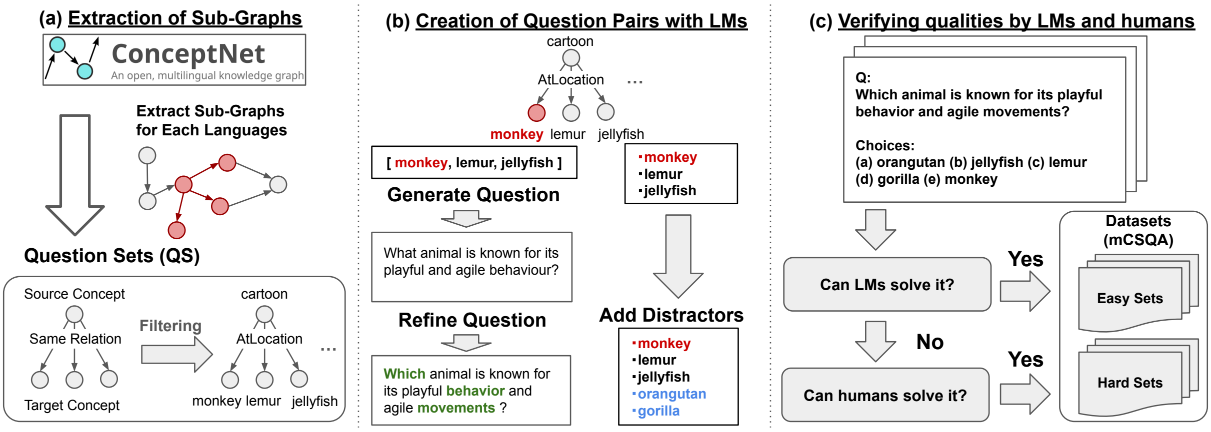

Figure 2: Creation process of mCSQA

We extract subgraphs from ConceptNet, as per Figure 2-(a), that have three distinct concept entities derived from queries of concept entities and relations for each language. CSQA uses only forward queries $(s, r, ?)$, but, similar to JCSQA, we also utilize backward queries $(?, r, t)$. We name this subgraph as Question Sets (QSs). After extraction, we filter the QSs like CSQA and JCSQA, and applies unified filtering in mCSQA as follows:

1. Similar to CSQA, we retain only QSs that contain any types of the 22 relations[2].
2. We filter out QSs where any of the concept entities consist of more than four words or only a single character[3].
3. We remove QSs where any pair of concept entities is connected by a 'Synonym' relation in ConceptNet, or where entities are substrings of each other.

After filtering with the above settings, we randomly selected 6,000 QSs for each language[4].

3.2 Create Questions with LMs

We employ the generative multilingual language model GPT-3.5[5] (Ouyang et al., 2022) to generate questions automatically to eliminate the human labor as done in CSQA and JCSQA.

[2]Antonym, AtLocation, CapableOf, Causes, CausesDesire, DefinedAs, DerivedFrom, Desires, DistinctFrom, EtymologicallyDerivedFrom, HasA, HasFirstSubevent, HasLastSubevent, HasPrerequisite, HasProperty, InstanceOf, MadeOf, MotivatedByGoal, NotDesires, PartOf, SymbolOf, UsedFor

[3]Unsegmented languages, like Japanese, are segmented by morphology in ConceptNet, so we can apply similar filtering.

[4]For French and Russian, the number of QSs did not reach 6,000, so we used all available QSs, totaling 4,125 and 3,901, respectively.

[5]We used gpt-3.5-turbo-1106.

step	temperature	top_p	seed
Creating question sentences	0.0	0.0	0
Refining question sentences	0.7	0.5	0
Adding additional distractors	1.2	0.7	0

Table 3: The hyper-parameters for each step

Our construction process comprises three steps of 'question generation', 'question refinement' and 'distractor augmentation' as shown in Figure 2-(b). Our step differs from CSQA in the refinement step since we need to improve the question generation from LM.

We designed prompts and tuned optimized hyper-parameters for each step for LMs. The details of the prompts are described in Appendix D, and the hyper-parameters are shown in Table 3.

Creating question sentences For each QS, we generated question sentences by LMs where, for each of the three target concept entities, only one serves as the answer. The prompt for LMs was inspired by the JCSQA filtering process for question creation in which systematic filtering uses textual information. The key instructions are as follows:

- Avoid including words of the target entities in the question sentence.
- Avoid using superficial information such as character count.
- End the sentence with a question mark (?).
- Be an objective question sentence.
- Consists of only one sentence.

After generating questions with LMs, we removed any question sentences that do not follow

	en	ja	zh	de	pt	nl	fr	ru
Total	14,722	15,695	17,254	16,542	16,679	15,992	10,770	10,215
Refined	3,654	12,007	6,534	765	585	7,927	3,109	6,734
pct. (%)	24.82	76.50	37.87	4.63	3.51	49.57	28.87	65.92

Table 4: The percentage of sentences refined

these instructions or contain inappropriate expressions through pattern matching[6].

Refining question sentences LMs do not always generate appropriate outputs resulting in unnatural expressions or degeneration (Liu et al., 2022c; Honovich et al., 2023; Raunak et al., 2023; Lin et al., 2020; Madaan et al., 2023). Hence, inspired by the idea of output refinement (Liu et al., 2022c; Raunak et al., 2023; Madaan et al., 2023), we refine unnatural generated question sentences into natural ones using the LM again and remove inappropriate questions as done in the previous step. Table 4 shows the percentage of sentence refinement.

Adding additional distractors We added additional incorrect choices to make the task more difficult as done in CSQA and JCSQA, but we leverage LM, not crowd workers, to formulate distractors that seemed plausible or related to the questions. Here, we asked LM to generate two plausible distractors given the three choices of a question without question itself in order to separate the question generation and answering capabilities of LMs. There is a risk of generating duplicated choices or adding correct choices since question sentence itself is not fed in this process. Hence, we remove such questions through manual verification in Section 3.3.

3.3 Question Quality Verification by LMs and Humans

In CSQA and JCSQA, every question is manually verified to remove low-quality questions, such as those with multiple correct answers or without correct answers in the choices. However, due to the large number of questions, manually verifying every question is not practical. Thus, we leverage simple active learning methodologies for annotation (Liu et al., 2022a; Bartolo et al., 2022; Li et al., 2023; Kratzwald et al., 2020). As shown in Figure 2-(c), initially, the LM verifies whether the questions can be answered or not, and only those questions that the LM cannot answer are manually verified.

[6]We detected inappropriate expressions using https://platform.openai.com/docs/guides/moderation.

	Train	Dev			Test		
	Total	Easy	Hard	Total	Easy	Hard	Total
English	10,910	1,071	292	1,363	1,071	292	1,363
Japanese	11,696	1,117	344	1,461	1,117	344	1,461
Chinese	12,159	972	546	1,518	972	546	1,518
German	12,504	1,279	283	1,562	1,279	283	1,562
Portuguese	12,659	1,234	348	1,582	1,234	348	1,582
Dutch	12,215	1,255	271	1,526	1,255	271	1,526
French	8,047	786	219	1,005	786	219	1,005
Russian	6,623	445	382	827	445	382	827

Table 5: The statistics of mCSQA

Verification by LMs The original questions can be categorized into three types: those questions 1) which are correctly answerable by LMs, 2) which are wrongly answered by LMs, but humans can choose the correct one, 3) which are not answerable either by LMs or humans due to flaws in the question. Therefore, first, we identify the set of questions LMs can answer, and then manually verify the questions that LMs could not answer correctly to remove flawed questions.

Verification by Humans We hired two crowd workers per language via Amazon Mechanical Turk (MTurk)[7]. The crowd workers were presented with the question sentence, choices, and answer, and they were asked to verify if the answer could be concluded from the question and choices. We retained only those questions on which all crowd workers agreed.

3.4 Data Splitting and Statistics

Similar to CSQA and JCSQA, we randomly split the data for each language into training, development, and test sets with an 80/10/10 split. The mCSQA is evaluated by accuracy following the standard practice in CSQA and JCSQA. Additionally, in Section 3.3, questions that LMs could answer correctly are categorized as Easy, and those answerable by human judgment are categorized as Hard for development and test sets.

Table 5 shows the number of questions per language and split, and Figure 3 shows the percentage filtered at each step. The total cost per question is 0.002 dollars for mCSQA compared to 0.33 dollars for CSQA, reducing the cost to less than one

[7]There are workers for each language on MTurk (Pavlick et al., 2014). We hired workers who have an approval rate greater than 90% with at least 50 approved HITs.

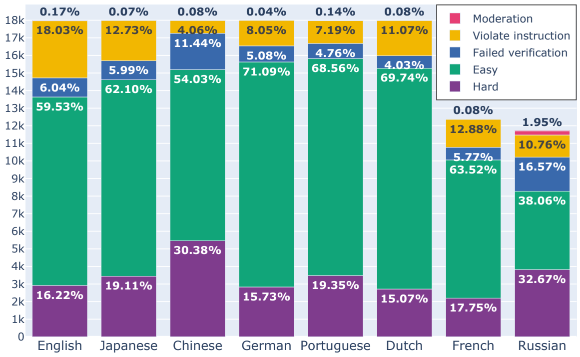

Figure 3: The percentage of sentences processed at each step. Easy and Hard were adopted for the dataset, while others were removed during the generation process.

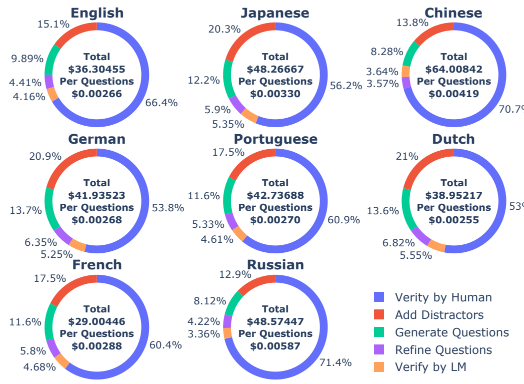

Figure 4: The cost details for each language and step.

hundredth. Figure 4 shows the detailed costs.

Appendix B discusses more detailed statistics, and Figure 11 shows examples of mCSQA.

4 Evaluation for mCSQA

We verify that the mCSQA dataset is meaningful for evaluating the common sense reasoning capability of LMs by using various multilingual LMs.

4.1 Experimental Setup

Settings for LMs We used mBERT (Devlin et al., 2019), XLM-100 (Conneau and Lample, 2019), XLM-R (Conneau et al., 2020), and mDeBERTa-v3 (He et al., 2023) as encoder-based multilingual LMs, Llama2-70B (Touvron et al., 2023), GPT-3.5 (Ouyang et al., 2022), and GPT-4 (OpenAI et al., 2024) as decoder-based multilingual LMs for the experiments. Decoder-based LMs inferred with 0-shot and 3-shot settings. For detailed experimental settings, please refer to the Appendix A.

Settings for human baseline We followed the CSQA setting and randomly selected 100 questions each from the validation and test data for every language to measure the human baseline. We hired five new crowd-workers per language on MTurk. The answers were decided by a majority vote for each question.

4.2 Evaluation Results

Table 6 shows the main results. Focusing on the performance of zero-shot setting of GPT-3.5, which was used for dataset creation, we find that its performance is equivalent to or worse than that of Encoder models like XLM-R_{LARGE} and mDeBERTa-v3 except for German and Russian. When comparing the results of GPT-3.5 with GPT-4, the performance of GPT-3.5 is inferior for most languages to that of GPT-4. This indicates that the questions GPT-3.5 failed to answer correctly are those that cannot be answered by the knowledge of GPT-3.5, and it implies that the root cause is a lack of knowledge or reasoning capability of GPT-3.5. Furthermore, focusing on Decoder-based models, the results are better in the 3-shot setting than in the 0-shot in most cases. This trend was observed even with the GPT-3.5 used for question creation.

The results show that the prompting technique is effective for mCSQA in exploiting the reasoning capabilities of decoder-based LMs. The trend is similar to other commonsense reasoning tasks like CSQA (Qin et al., 2023; Chowdhery et al., 2023; Wei et al., 2022c; Brown et al., 2020; Dou and Peng, 2022), indicating that mCSQA can be equally effective as a dataset for commonsense reasoning tasks. Finally, when compared to the human baseline, there is a significant gap in the results of all LMs. Thus, it can be said that even when using LMs for question creation, it is possible to create a dataset with sufficient quality and difficulty for the LMs themselves.

5 Discussion

5.1 Comparison of Easy vs. Hard

We compare the accuracy of Easy and Hard sets for more fine-grained analysis. Figure 5 shows the results in the test split. GPT-3.5 and GPT-4 could choose the answer correctly in most cases for the Easy sets, but the accuracy is lower in the Hard sets with a significant gap when compared with human results; note that GPT-3.5 cannot answer there sets during the dataset creation. The other LMs also show a gap in evaluation accuracy with results for Hard sets being lower than those for Easy ones.

These results, specifically the trend observed

	English		Japanese		Chinese		German		Portuguese		Dutch		French		Russian	
	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test
Human (Rand. 100 sent.)	87.0	93.0	89.0	95.0	91.0	87.0	96.0	96.0	93.0	93.0	98.0	97.0	96.0	92.0	87.0	94.0
mBERT-cased	60.6	61.3	66.0	63.5	65.9	63.5	58.6	57.9	65.2	61.5	54.8	57.8	46.3	47.3	32.2	31.3
mBERT-uncased	63.4	65.2	61.3	58.9	64.0	62.0	59.3	60.3	67.6	63.9	57.3	56.9	51.1	52.4	32.5	34.0
XLM-100	57.2	59.0	60.2	58.8	60.0	61.5	54.4	54.7	62.7	59.5	52.2	52.0	35.3	35.0	23.2	26.0
XLM-R_{BASE}	68.0	69.1	68.5	66.2	69.8	68.3	63.9	62.8	69.5	67.3	62.0	64.0	47.6	45.5	36.9	37.0
XLM-R_{LARGE}	77.2	77.5	75.7	72.6	**75.0**	**74.1**	76.2	75.4	79.0	76.4	73.0	74.7	62.0	62.3	48.9	49.5
mDeBERTa-v3	76.6	79.2	77.2	74.1	74.6	72.0	75.7	77.5	78.3	78.2	72.7	74.9	62.1	62.4	51.3	49.9
Llama2-70B (0-shot)	48.1	47.7	25.6	24.8	26.5	25.9	32.5	32.7	38.7	37.6	40.9	39.4	42.3	44.1	23.5	22.9
Llama2-70B (3-shot)	57.1	55.5	47.4	46.6	33.3	30.2	63.1	62.9	65.0	63.7	60.8	62.3	57.8	56.7	30.8	32.3
GPT-3.5 (0-shot)	76.7	77.0	76.3	76.7	64.0	63.6	81.3	81.4	77.9	77.7	82.1	81.5	78.6	77.1	53.3	**53.0**
GPT-3.5 (3-shot)	77.2	78.4	77.5	77.0	65.3	64.3	**83.2**	81.4	78.5	78.0	81.8	80.5	78.4	76.5	**54.1**	50.1
GPT-4 (0-shot)	**80.9**	80.9	78.4	77.2	66.0	65.6	81.0	81.0	78.6	77.6	**83.4**	81.5	78.8	77.0	49.9	47.8
GPT-4 (3-shot)	80.5	**81.0**	**78.5**	**77.5**	67.2	66.9	82.6	**81.6**	**80.5**	**78.8**	83.3	**81.6**	**79.0**	**77.4**	50.1	48.9

Table 6: The results on mCSQA (acc. %)

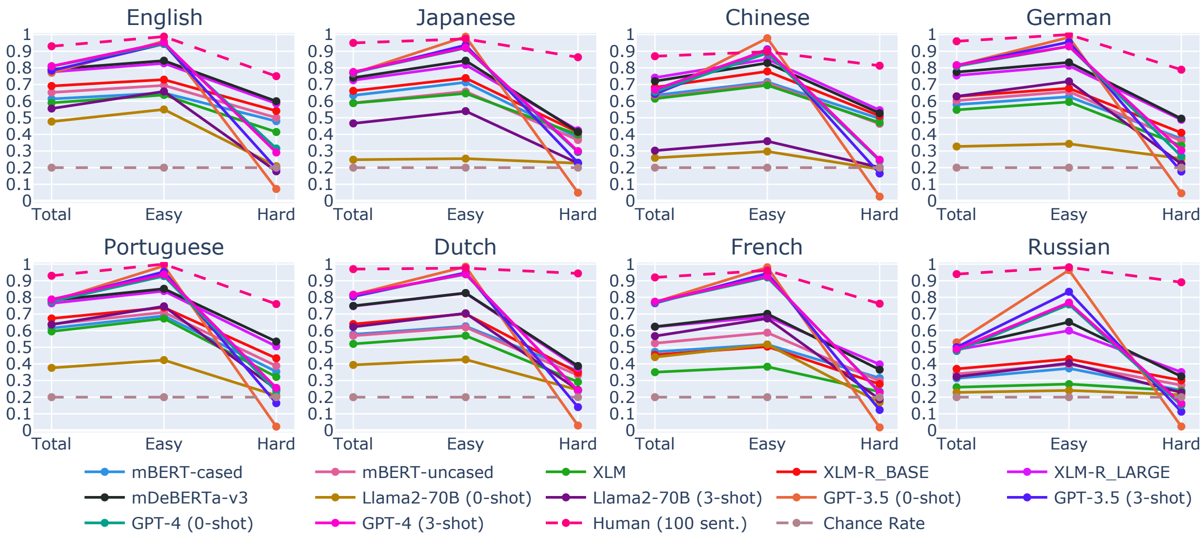

Figure 5: Comparison of the evaluation accuracy between Easy and Hard sets.

with GPT-3.5, show that even if LMs can create questions, it does not necessarily mean that they can answer them, and it entails that the question creation and answering are totally different capabilities. Therefore, we conclude that LMs can substitute for humans in parts of dataset creation processes from structured data and common sense reasoning task creations.

5.2 Evaluation of Multilingual LMs' Cross-Lingual Transfer Capabilities

The cross-lingual transfer performance of multilingual LMs is often evaluated from English to other language directions due to linguistic resource reasons. The X-CSQA dataset (Lin et al., 2021), which consists solely of machine-translated questions from CSQA's development and test splits, captures only the one-way cross-lingual transfer performance of LMs that were trained in English to evaluate their performance in other languages. In contrast, mCSQA supports the evaluation of cross-lingual language transfer performance in any directions among multilingual LMs that were trained in each of the eight languages.

Figure 6 shows the results of the multilingual LM, XLM-R_{LARGE}, which was fine-tuned in each of the eight languages separately and then evaluated across all eight languages on mCSQA, using the same settings as in Table 10. The results from Figure 6 show that, regardless of the language in which they were trained, cross-lingual transfer abilities are observed in most cases for any languages given the relative lower drop of performance when compared with the monolingual performance. Moreover, in the Easy sets, the drop is within 10% for most language pairs, while in the Hard sets, it ex-

Figure 6: The language transfer performance of XLM-R$_{\text{LARGE}}$. The y-axis indicates the languages in which the model was fine-tuned, while the x-axis indicates the languages used for evaluation. It shows the percentage of performance achieved when compared with the model trained and evaluated in the same language.

		Test (%)							
		en	ja	zh	de	pt	nl	fr	ru
Total	Mono.	77.5	72.6	74.1	75.4	76.4	74.7	62.3	49.5
	Multi.	81.4	74.6	74.2	77.8	79.9	77.0	65.7	54.2
	Δ	3.9	2.0	0.1	2.4	3.5	2.3	2.4	4.7
	Unseen	80.0	71.8	71.3	76.6	76.0	76.6	64.9	51.8
Easy	Mono.	82.8	81.9	85.3	81.2	83.8	82.5	68.4	60.0
	Multi.	86.5	83.4	85.7	84.4	86.6	84.9	73.2	70.1
	Δ	3.7	2.5	0.4	3.2	2.8	2.4	4.8	10.1
	Unseen	85.4	81.2	84.1	83.4	83.1	84.1	71.8	68.8
Hard	Mono.	58.6	42.4	54.6	49.1	50.6	37.6	39.7	35.1
	Multi.	62.7	45.9	53.7	47.7	56.0	40.2	38.8	35.6
	Δ	4.1	3.5	-0.9	-1.4	5.4	3.6	-0.9	0.5
	Unseen	60.3	41.0	48.7	45.9	51.2	39.1	38.6	31.7

Table 7: The performance comparison of XLM-R$_{\text{LARGE}}$ on test data for each language when trained on monolingual training data versus multilingual data. Δ means the differences in performance between the two settings. Unseen means the accuracy when trained on all training data except for the evaluation language.

ceeds 20%. This indicates that questions that are relatively easy to judge (Easy sets) facilitate the language transfer capability, but questions requiring deep background knowledge (Hard sets) necessitate language-specific training and the development of LMs.

5.3 Which is Better: Monolingual Fine-tuning or Multilingual Fine-tuning?

Some studies (Tran and Bisazza, 2019; Dhamecha et al., 2021; Trotta et al., 2021; Barbieri et al., 2022; Portelli et al., 2023) reported that multilingual fine-tuning could improve a part of NLU task performance more than monolingual tuning alone. On the other hand, several studies (Tsai et al., 2019; Kondratyuk, 2019; Rønningstad, 2023; Kondratyuk and Straka, 2019) reported that it did not always improve performance in some tasks. We analyzed whether multilingual fine-tuning is effective for commonsense reasoning tasks through mCSQA. We used the whole shuffled training split data in all languages and fine-tuned XLM-R$_{\text{LARGE}}$ with the same setting as in Table 10. Table 7 compares the accuracy between monolingual fine-tuning, where tuning and evaluation are in the same language, and multilingual fine-tuning, where tuning is performed for all languages, evaluated for each language's accuracy score. These results show that most languages observed improvements, especially in all cases in Easy sets. However, in Hard sets, some cases observed a decline in performance compared to the monolingual setting. Therefore, while training in a multilingual setting generally promotes accuracy improvement, multilingual training might lead to the loss of language-specific commonsense information for questions requiring more human commonsense. This analysis complements the previous reports (Dhamecha et al., 2021; Zhang et al., 2023a; Hu et al., 2021; Mueller et al., 2020) on the successes and failures of multilingual training.

Furthermore, Table 7 shows the evaluation results of cross-lingual performance in the unseen setting, where the model was not trained on the language for evaluation data. While some languages outperform the monolingual setting, overall results indicate that training with target language data consistently yields better outcomes. This suggests that target language data acts as the secret sauce for enhancing NLU performance. Therefore, it suggests that for language-specific deep knowledge and cultural understanding, language-transfer capability alone is insufficient, and training with datasets focused on language-specific knowledge is necessary.

5.4 Case Study for Improvement through Few-Shot Learning

As shown in Figure 5, GPT-3.5 correctly answers most questions in the Easy setting of mCSQA, but in the Hard setting, it fails to answer most questions in the 0-shot setting. This is because GPT-3.5 is used for quality filtering of mCSQA in Section 3.3, making it inherently unable to answer the questions in the Hard setting in the 0-shot setting. However,

Question	Answer	0-shot	3-shot
Which types of aquatic animals are commonly found in the open sea? (a) marine life, (b) earth, (c) waves, (d) coastline, (e) oceanic fish	e	a	e
What is the purpose of using hand gestures while driving? (a) determine what caused noise, (b) giving signal to, (c) checking for any potential dangers, (d) warning, (e) investigating the source of the noise	b	c	b

Table 8: Examples of GPT-3.5 correctly answering in a 3-shot setting. In the top example, a 0-shot setting would choose "marine life", but considering the phrase "in the open sea" in the question, the answer should be narrowed down to "oceanic fish". On the other hand, in the bottom example, it chooses "checking for any potential dangers", but "hand gestures while driving" can include broader, non-dangerous signals such as thank-you gestures. Therefore, the broader "giving signal to" is correct. In this way, the 3-shot setting tended to allow for appropriately granular answers that matched the intent of the question.

in the 3-shot setting, it shows improvement for some questions. Table 8 shows examples of questions correctly answered in the 3-shot setting. Both examples in Table 8 are mainly due to the granularity of the answers. The 3-shot setting promotes answers at an appropriate granularity for questions that are difficult to judge due to inclusive relationships.

In the top example in Table 8, careful reading of the questions narrows down the answer choices. On the other hand, in the bottom example, considering various common knowledge in daily life helps to choose the most appropriate answer. Similar characteristics were observed for other languages as well. For more details, qualitative analyses of the mCSQA dataset are described in Appendix B.

6 Conclusion and Future Directions

We proposed an efficient and low-cost method for creating NLU datasets from structured data by utilizing generative LMs as an alternative to traditional human annotation, often crowdsourced. Inspired by CSQA and JCSQA, we created the multilingual commonsense reasoning task dataset, mCSQA, using GPT-3.5 from the structured multilingual knowledge base ConceptNet. We demonstrated that mCSQA is useful for evaluating the commonsense reasoning capabilities of LMs. We also analyzed the language-transfer capability beyond English with mCSQA and examined the language-specific learning from two aspects: question difficulty and language information. Moreover, our study has shown that the use of multilingual LMs enables the construction of multilingual datasets. Therefore, our method can significantly reduce human labor and financial costs.

In this study, we used a single multilingual LM, but since each step is independent, it is possible to replace the LM used in each step with another one. Furthermore, each step can be applied modularly to other methods, making it possible to use this method for creating multilingual datasets, such as those expanded through translation and manual refinement (Yanaka and Mineshima, 2022; Seo et al., 2022). We aim to extend this method to other types of commonsense reasoning tasks and NLU tasks, to efficiently create multilingual data and conduct a more comprehensive analysis of transfer capabilities across a broader range of tasks and languages.

We focused on language-specific commonsense, but languages are shared across various regions. For example, English is spoken in the United States, the United Kingdom, India, Australia, and many other regions each of which is geographically distant and diverse in terms of climate, food, and culture. Therefore, it will be necessary to create more detailed commonsense tasks that consider cultural differences rather than just language such as Kabra et al. (2023); Khanuja et al. (2024); Kim et al. (2024); Cao et al. (2024); Fung et al. (2024); Lee et al. (2024); Shwartz (2022); Hovy and Yang (2021); Yin et al. (2022); Shi et al. (2024). Our dataset construction method can be useful in creating various commonsense reasoning datasets that outgrow language limits.

7 Limitations

Data Resources The number of multilingual resources is significantly smaller than that of monolingual resources. Additionally, quality is not consistent, and there are imbalances in data volume across languages in these multilingual resources. In this study, we used ConceptNET, a multilingual knowledge base, and encountered these issues as well. For example, despite Spanish having a significantly higher entity count, it obtained fewer QSs due to its inability to meet the required conditions because of ConceptNet's sparsity issue, and thus it was excluded from the language selection for mCSQA. We believe these problems can be addressed through the automatic generation of knowledge bases (Zhang et al., 2020b,a; West et al., 2022; Ide et al., 2023; Nguyen et al., 2023) and data augmentation techniques for knowledge bases (Malaviya et al., 2020; Ju et al., 2022; Wu et al., 2023; Shen et al., 2023), supported by their pre-trained knowledge (Sakai et al., 2023).

Dataset Quality In this study, we used GPT-3.5 and simple prompts for data creation. Therefore, there is room for improvement in the selection of LMs and the refinement of prompts. In a pilot study, we tried using GPT-4 and recognized that it is more capable of creating datasets. However, due to budgetary constraints, we have used GPT-3.5 in this study. Thus, it may become possible to create higher quality datasets at a lower cost when the API prices decrease or by switching to other strong LMs such as Gemini (Team et al., 2023), Mixtral (Jiang et al., 2024), Llama (Touvron et al., 2023), phi (Abdin et al., 2024) or Qwen (Bai et al., 2023). Additionally, employing prompt strategies that leverage the capabilities of LMs, such as Chain of Thought (CoT) (Wei et al., 2022c), Tree of Thought (ToT) (Yao et al., 2023a) and ReAct (Yao et al., 2023b), could potentially lead to the production of higher quality datasets.

Verification of dataset quality by humans The human baselines decreased the evaluation result under the Hard sets compared to Easy sets in Figure 5. Therefore, there exists a risk that the Hard sets include flawed questions, even after manual quality verification. The JCSQA has pointed out that such low-quality questions are included in CSQA, and we have confirmed that they are similarly present in JCSQA. Thus, it is extremely difficult to completely eliminate such low-quality questions. Comparing the percentage of data removed in quality verification, CSQA is 25% (3995/16242), and JCSQA is 19% (2643/13906), whereas for mCSQA, it is 27% (604/2226) and 23% (599/2510) respectively, according to Figure 3 and referenced in their respective papers. This indicates that the filtering ratios are almost comparable when compared to those, showing that this is not a problem unique to mCSQA. Therefore, reinforcing quality verification to filter out low-quality questions is a challenge in our future studies. However, as Figure 4 shows, since more than half of the costs are already spent on manual quality verification, simply hiring more crowd workers would not be a better choice. Hence, exploring more efficient methods of quality verification as an alternative or to assist crowd workers in the future is necessary.

Human baseline The experimental results include human baselines using small sets of samples. However, Tedeschi et al. (2023) argue that human baselines may lack reliability due to factors such as the payment issues for crowd workers and the impact of random samples. Therefore, it should be noted that the human baselines in this study are merely reference values.

8 Ethical Considerations

License The mCSQA dataset was created entirely from the outputs of GPT-3.5 and is therefore subject to OpenAI's license terms[8]. OpenAI assigns to us all rights, title, and interest in and to the output. As a result, we are retaining the ownership rights. There are no restrictions on distributing the datasets, but using OpenAI's model output to develop models that compete with OpenAI is prohibited. However, it's possible that these terms may change, and there may be a need to impose distribution restrictions depending on the terms.

Moderation We eliminated potentially harmful questions such as violence, sexual content, and hate speech by screening through OpenAI moderation APIs[9]. However, in the commonsense reasoning dataset, it cannot be guaranteed that it does not include questions that contain societal biases as collective knowledge. This issue has also been pointed out in existing datasets such as CSQA, JCSQA, and other commonsense reasoning datasets, and it is challenging to determine what is considered commonsense constitutes bias (Rajani et al., 2019; Sap et al., 2020; Bauer et al., 2023; An et al., 2023). If you encounter any harmful questions that contain such biases, please report them.

Translation Tool We used DeepL Pro[10] to translate the example sentence, especially Table 1, to avoid arbitrary translation. The copyright of the translation sentences belongs to us[11].

[8] https://openai.com/policies/terms-of-use
[9] https://platform.openai.com/docs/guides/moderation
[10] https://www.deepl.com/translator
[11] https://www.deepl.com/pro-license

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone.

Anurag Acharya, Kartik Talamadupula, and Mark A Finlayson. 2020. Towards an atlas of cultural commonsense for machine reasoning.

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, and Joyce Nakatumba-Nabende. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for CommonsenseQA: New Dataset and Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065, Online. Association for Computational Linguistics.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. 2022. Do as i can, not as i say: Grounding language in robotic affordances.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Haozhe An, Zongxia Li, Jieyu Zhao, and Rachel Rudinger. 2023. SODAPOP: Open-ended discovery of social biases in social commonsense reasoning models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1573–1596, Dubrovnik, Croatia. Association for Computational Linguistics.

Jordi Armengol-Estapé, Ona de Gibert Bonet, and Maite Melero. 2022. On the multilingual capabilities of very large-scale English language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3056–3068, Marseille, France. European Language Resources Association.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Abhijeet Awasthi, Nitish Gupta, Bidisha Samanta, Shachi Dave, Sunita Sarawagi, and Partha Talukdar. 2023. Bootstrapping multilingual semantic parsers using large language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2455–2467, Dubrovnik, Croatia. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Max Bartolo, Tristan Thrush, Sebastian Riedel, Pontus Stenetorp, Robin Jia, and Douwe Kiela. 2022. Models in the loop: Aiding crowdworkers with generative annotation assistants. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3754–3767, Seattle, United States. Association for Computational Linguistics.

Lisa Bauer, Hanna Tischer, and Mohit Bansal. 2023. Social commonsense for explanation and cultural bias discovery. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3745–3760, Dubrovnik, Croatia. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York City. Association for Computational Linguistics.

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024. Cultural Adaptation of Recipes. *Transactions of the Association for Computational Linguistics*, 12:80–99.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Du Chen, Yi Huang, Xiaopu Li, Yongqiang Li, Yongqiang Liu, Haihui Pan, Leichao Xu, Dacheng Zhang, Zhipeng Zhang, and Kun Han. 2024. Orion-14b: Open-source multilingual large language models.

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.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

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*, volume 32. Curran Associates, Inc.

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.

Steven Coyne, Keisuke Sakaguchi, Diana Galvan-Sosa, Michael Zock, and Kentaro Inui. 2023. Analyzing the performance of gpt-3.5 and gpt-4 in grammatical error correction.

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.

Tejas Dhamecha, Rudra Murthy, Samarth Bharadwaj, Karthik Sankaranarayanan, and Pushpak Bhattacharyya. 2021. Role of Language Relatedness in Multilingual Fine-tuning of Language Models: A Case Study in Indo-Aryan Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8584–8595, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zi-Yi Dou and Nanyun Peng. 2022. Zero-shot commonsense question answering with cloze translation and consistency optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10572–10580.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tao Fang, Shu Yang, Kaixin Lan, Derek F. Wong, Jinpeng Hu, Lidia S. Chao, and Yue Zhang. 2023. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation.

Anjalie Field and Yulia Tsvetkov. 2020. Unsupervised discovery of implicit gender bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 596–608, Online. Association for Computational Linguistics.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition & lm benchmarking.

Himanshu Gupta, Kevin Scaria, Ujjwala Anantheswaran, Shreyas Verma, Mihir Parmar, Saurabh Arjun Sawant, Chitta Baral, and Swaroop Mishra. 2023. Targen: Targeted data generation with large language models.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023. Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214, Toronto, Canada. Association for Computational Linguistics.

Yuta Hayashibe. 2020. Japanese realistic textual entailment corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6827–6834, Marseille, France. European Language Resources Association.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. 2022. Generate, annotate, and learn: NLP with synthetic text. *Transactions of the Association for Computational Linguistics*, 10:826–842.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada. Association for Computational Linguistics.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Hai Hu, Ziyin Zhang, Weifang Huang, Jackie Yan-Ki Lai, Aini Li, Yina Patterson, Jiahui Huang, Peng Zhang, Chien-Jer Charles Lin, and Rui Wang. 2023. Revisiting acceptability judgements.

Hai Hu, He Zhou, Zuoyu Tian, Yiwen Zhang, Yina Patterson, Yanting Li, Yixin Nie, and Kyle Richardson. 2021. Investigating transfer learning in multilingual pre-trained language models through Chinese natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3770–3785, Online. Association for Computational Linguistics.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual 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 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Tenghao Huang, Faeze Brahman, Vered Shwartz, and Snigdha Chaturvedi. 2021. Uncovering implicit gender bias in narratives through commonsense inference. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3866–3873, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J. Paul. 2020. Multilingual Twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1440–1448, Marseille, France. European Language Resources Association.

Tatsuya Ide, Eiki Murata, Daisuke Kawahara, Takato Yamazaki, Shengzhe Li, Kenta Shinzato, and Toshinori Sato. 2023. Phalm: Building a knowledge graph from scratch by prompting humans and a language model.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2023. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries.

Jinhao Ju, Deqing Yang, and Jingping Liu. 2022. Commonsense knowledge base completion with relational graph attention network and pre-trained language model. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 4104–4108, New York, NY, USA. Association for Computing Machinery.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284, Toronto, Canada. Association for Computational Linguistics.

Masahiro Kaneko and Naoaki Okazaki. 2023. Reducing sequence length by predicting edit spans with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10017–10029, Singapore. Association for Computational Linguistics.

Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. 2023. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations.

Pride Kavumba, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reisert, and Kentaro Inui. 2019. When choosing plausible alternatives, clever hans can be clever. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 33–42, Hong Kong, China. Association for Computational Linguistics.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

Tannon Kew, Florian Schottmann, and Rico Sennrich. 2023. Turning english-centric llms into polyglots: How much multilinguality is needed?

Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. 2024. An image speaks a thousand words, but can everyone listen? on image transcreation for cultural relevance.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. CLIcK: A benchmark dataset of cultural and linguistic intelligence in Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346, Torino, Italia. ELRA and ICCL.

Vid Kocijan, Ernest Davis, Thomas Lukasiewicz, Gary Marcus, and Leora Morgenstern. 2023. The defeat of the winograd schema challenge.

Dan Kondratyuk. 2019. Cross-lingual lemmatization and morphology tagging with two-stage multilingual BERT fine-tuning. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 12–18, Florence, Italy. Association for Computational Linguistics.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. 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 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Fajri Koto, Timothy Baldwin, and Jey Han Lau. 2022. Cloze evaluation for deeper understanding of commonsense stories in Indonesian. In *Proceedings of the First Workshop on Commonsense Representation and Reasoning (CSRR 2022)*, pages 8–16, Dublin, Ireland. Association for Computational Linguistics.

Bernhard Kratzwald, Stefan Feuerriegel, and Huan Sun. 2020. Learning a Cost-Effective Annotation Policy for Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3051–3062, Online. Association for Computational Linguistics.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Varun Kumar, Hadrien Glaude, Cyprien de Lichy, and Wlliam Campbell. 2019. A closer look at feature space data augmentation for few-shot intent classification. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 1–10, Hong Kong, China. Association for Computational Linguistics.

Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. 2022. JGLUE: Japanese general language understanding evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2957–2966, Marseille, France. European Language Resources Association.

Sang Kwon, Gagan Bhatia, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Beyond English: Evaluating LLMs for Arabic grammatical error correction. In *Proceedings of ArabicNLP 2023*, pages 101–119, Singapore (Hybrid). Association for Computational Linguistics.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Kenton Lee, Kelvin Guu, Luheng He, Tim Dozat, and Hyung Won Chung. 2021. Neural data augmentation via example extrapolation.

Nayeon Lee, Yejin Bang, Holy Lovenia, Samuel Cahyawijaya, Wenliang Dai, and Pascale Fung. 2023. Survey of social bias in vision-language models.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. Exploring cross-cultural differences in english hate speech annotations: From dataset construction to analysis.

Heather Lent and Anders Søgaard. 2021. Common sense bias in semantic role labeling. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 114–119, Online. Association for Computational Linguistics.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy Chen, Zhengyuan Liu, and Diyi Yang. 2023. CoAnnotating: Uncertainty-guided work allocation between human and large language models for data annotation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1487–1505, Singapore. Association for Computational Linguistics.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1274–1287, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022a. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022b. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

Xin Liu, Dayiheng Liu, Baosong Yang, Haibo Zhang, Junwei Ding, Wenqing Yao, Weihua Luo, Haiying Zhang, and Jinsong Su. 2022c. KGR4: Retrieval, retrospect, refine and rethink for commonsense generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11029–11037.

Mengsay Loem, Masahiro Kaneko, Sho Takase, and Naoaki Okazaki. 2023. Exploring effectiveness of GPT-3 in grammatical error correction: A study on performance and controllability in prompt-based methods. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 205–219, Toronto, Canada. Association for Computational Linguistics.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Chaitanya Malaviya, Chandra Bhagavatula, Antoine Bosselut, and Yejin Choi. 2020. Commonsense knowledge base completion with structural and semantic context. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(03):2925–2933.

Shervin Malmasi and Mark Dras. 2015. Large-scale native language identification with cross-corpus evaluation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1403–1409, Denver, Colorado. Association for Computational Linguistics.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. In *Advances in Neural Information Processing Systems*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: GeneraLized and COntextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4569–4586, Online. Association for Computational Linguistics.

David Mueller, Nicholas Andrews, and Mark Dredze. 2020. Sources of transfer in multilingual named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8093–8104, Online. Association for Computational Linguistics.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 1907–1917, New York, NY, USA. Association for Computing Machinery.

Kazumasa Omura, Daisuke Kawahara, and Sadao Kurohashi. 2020. A method for building a commonsense inference dataset based on basic events. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2450–2460, Online. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Simon Ostermann, Ashutosh Modi, Michael Roth, Stefan Thater, and Manfred Pinkal. 2018. MCScript: A novel dataset for assessing machine comprehension using script knowledge. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo

Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. 2021. KLUE: Korean language understanding evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The language demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2:79–92.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4.

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.

Beatrice Portelli, Alessandro Tremamunno, Simone Scaboro, Emmanuele Chersoni, and Giuseppe Serra. 2023. Ailabud at the ntcir-17 mednlp-sc task: Monolingual vs multilingual fine-tuning for ade classification. NII Institutional Repository.

Rifki Afina Putri, Faiz Ghifari Haznitrama, Dea Adhista, and Alice Oh. 2024. Can llm generate culturally relevant commonsense qa data? case study in indonesian and sundanese.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore. Association for Computational Linguistics.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Awadalla, and Arul Menezes. 2023. Leveraging GPT-4 for automatic translation post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12009–12024, Singapore. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Egil Rønningstad. 2023. UIO at SemEval-2023 task 12: Multilingual fine-tuning for sentiment classification in low-resource languages. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1054–1060, Toronto, Canada. Association for Computational Linguistics.

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. 2020. LAReQA: Language-agnostic answer retrieval from a multilingual pool. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5919–5930, Online. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2023. Does pre-trained language model actually infer unseen links in knowledge graph completion? *CoRR*, abs/2311.09109.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019a. Atomic: an atlas of machine commonsense for if-then reasoning. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–

6951, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tim Schott, Daniel Furman, and Shreshta Bhat. 2023. Polyglot or not? measuring multilingual encyclopedic knowledge in foundation models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11238–11253, Singapore. Association for Computational Linguistics.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jaehyung Seo, Seounghoon Lee, Chanjun Park, Yoonna Jang, Hyeonseok Moon, Sugyeong Eo, Seonmin Koo, and Heuiseok Lim. 2022. A dog is passing over the jet? a text-generation dataset for Korean commonsense reasoning and evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2233–2249, Seattle, United States. Association for Computational Linguistics.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Synthetic prompting: generating chain-of-thought demonstrations for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. RussianSuperGLUE: A Russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726, Online. Association for Computational Linguistics.

Xiangqing Shen, Siwei Wu, and Rui Xia. 2023. Dense-ATOMIC: Towards densely-connected ATOMIC with high knowledge coverage and massive multi-hop paths. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13292–13305, Toronto, Canada. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Chunhua yu, Raya Horesh, Rogério Abreu de Paula, and Diyi Yang. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies.

Vered Shwartz. 2022. Good night at 4 pm?! time expressions in different cultures. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2842–2853, Dublin, Ireland. Association for Computational Linguistics.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning.

Denis Smirnov. 2019. Neural network-based models with commonsense knowledge for machine reading comprehension. In *Proceedings of the Student Research Workshop Associated with RANLP 2019*, pages 90–94, Varna, Bulgaria. INCOMA Ltd.

Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. Evaluation metrics in the era of GPT-4: Reliably evaluating large language models on sequence to sequence tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8776–8788, Singapore. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Ieva Staliūnaitė, Philip John Gorinski, and Ignacio Iacobacci. 2021. Improving commonsense causal reasoning by adversarial training and data augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13834–13842.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Daniel Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Norio Takahashi, Tomohide Shibata, Daisuke Kawahara, and Sadao Kurohashi. 2019. Machine comprehension improves domain-specific Japanese predicate-argument structure analysis. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 98–104, Hong Kong, China. Association for Computational Linguistics.

Ekaterina Taktasheva, Alena Fenogenova, Denis Shevelev, Nadezhda Katricheva, Maria Tikhonova, Albina Akhmetgareeva, Oleg Zinkevich, Anastasiia

Bashmakova, Svetlana Iordanskaia, Valentina Kurenshchikova, Alena Spiridonova, Ekaterina Artemova, Tatiana Shavrina, and Vladislav Mikhailov. 2022. TAPE: Assessing few-shot Russian language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2472–2497, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. CommonsenseQA 2.0: Exposing the limits of AI through gamification. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Alexandre Tamborrino, Nicola Pellicanò, Baptiste Pannier, Pascal Voitot, and Louise Naudin. 2020. Pre-training is (almost) all you need: An application to commonsense reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3878–3887, Online. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Martin Chadwick, Gaurav Singh Tomar, Xavier Garcia, Evan Senter, Emanuel Taropa, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Yujing Zhang, Ravi Addanki, Antoine Miech, Annie Louis, Laurent El Shafey, Denis Teplyashin, Geoff Brown, Elliot Catt, Nithya Attaluri, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaly Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, Hanzhao Lin, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yong Cheng, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat,

Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, YaGuang Li, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Gamaleldin Elsayed, Ed Chi, Mahdis Mahdieh, Ian Tenney, Nan Hua, Ivan Petrychenko, Patrick Kane, Dylan Scandinaro, Rishub Jain, Jonathan Uesato, Romina Datta, Adam Sadovsky, Oskar Bunyan, Dominik Rabiej, Shimu Wu, John Zhang, Gautam Vasudevan, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Betty Chan, Pam G Rabinovitch, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Sahitya Potluri, Jane Park, Elnaz Davoodi, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Chris Gorgolewski, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Paul Suganthan, Evan Palmer, Geoffrey Irving, Edward Loper, Manaal Faruqui, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Michael Fink, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marin Georgiev, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Minnie Lui, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Lam Nguyen Thiet, Daniel Andor, Pedro Valenzuela, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Sarmishta Velury, Sebastian Krause, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Tejasi Latkar, Mingyang Zhang, Quoc Le, Elena Allica Abellan, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Sid Lall, Ken Franko, Egor Filonov, Anna Bulanova, Rémi Leblond, Vikas Yadav, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Hao Zhou, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Jeremiah Liu, Mark Omernick, Colton Bishop, Chintu Kumar, Rachel Sterneck, Ryan Foley, Rohan Jain, Swaroop Mishra, Jiawei Xia, Taylor Bos, Geoffrey Cideron, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Petru Gurita, Hila Noga, Premal Shah, Daniel J. Mankowitz, Alex Polozov, Nate Kushman, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Anhad Mohananey, Matthieu Geist, Sidharth Mudgal, Sertan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Quan Yuan, Sumit Bagri, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Aliaksei Severyn, Jonathan Lai, Kathy Wu, Heng-Tze Cheng, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Mark Geller, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Andrei Sozanschi, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Tal-

bert, Abhimanyu Goyal, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Sabaer Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Tao Zhu, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Dustin Tran, Yeqing Li, Nir Levine, Ariel Stolovich, Norbert Kalb, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Balaji Lakshminarayanan, Charlie Deck, Shyam Upadhyay, Hyo Lee, Mike Dusenberry, Zonglin Li, Xuezhi Wang, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Summer Yue, Sho Arora, Eric Malmi, Daniil Mirylenka, Qijun Tan, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Steven Zheng, Francesco Pongetti, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Ragha Kotikalapudi, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Chenkai Kuang, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Pei Sun, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Ishita Dasgupta, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Yuan Liu, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Ivo Penchev, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Adam Kurzrok, Lynette Webb, Sahil Dua, Dong Li, Preethi Lahoti, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Taylan Bilal, Evgenii Eltyshev, Daniel Balle, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Adams Yu, Christof Angermueller, Xiaowei Li, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Kevin Brooks, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Komal Jalan, Dinghua Li, Ginger Perng, Blake Hechtman, Parker Schuh, Milad Nasr, Mia Chen, Kieran Milan, Vladimir Mikulik, Trevor Strohman, Juliana Franco, Tim Green, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2023. Gemini: A family of highly capable multimodal models.

Simone Tedeschi, Johan Bos, Thierry Declerck, Jan Hajič, Daniel Hershcovich, Eduard Hovy, Alexander Koller, Simon Krek, Steven Schockaert, Rico Sennrich, Ekaterina Shutova, and Roberto Navigli. 2023. What's the meaning of superhuman performance in today's NLU? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12471–12491, Toronto, Canada. Association for Computational Linguistics.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Ke Tran and Arianna Bisazza. 2019. Zero-shot dependency parsing with pre-trained multilingual sentence representations. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 281–288, Hong Kong, China. Association for Computational Linguistics.

Daniela Trotta, Raffaele Guarasci, Elisa Leonardelli, and Sara Tonelli. 2021. Monolingual and cross-lingual acceptability judgments with the Italian CoLA corpus. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2929–2940, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Henry Tsai, Jason Riesa, Melvin Johnson, Naveen Arivazhagan, Xin Li, and Amelia Archer. 2019. Small and practical BERT models for sequence labeling. 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 3632–3636, Hong Kong, China. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Chenhao Wang, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. 2022. CN-AutoMIC: Distilling Chinese commonsense knowledge from pretrained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9253–9265, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021a. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Zirui Wang, Adams Wei Yu, Orhan Firat, and Yuan Cao. 2021b. Towards zero-label language learning.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022c. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.

Chenxi Whitehouse, Monojit Choudhury, and Alham Aji. 2023. LLM-powered data augmentation for enhanced cross-lingual performance. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 671–686, Singapore. Association for Computational Linguistics.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

Brandon T Willard and Rémi Louf. 2023. Efficient guided generation for llms. *arXiv preprint arXiv:2307.09702*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Siwei Wu, Xiangqing Shen, and Rui Xia. 2023. Commonsense knowledge graph completion via contrastive pretraining and node clustering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13977–13989, Toronto, Canada. Association for Computational Linguistics.

Canwen Xu, Wangchunshu Zhou, Tao Ge, Ke Xu, Julian McAuley, and Furu Wei. 2021. Blow the dog whistle: A Chinese dataset for cant understanding with common sense and world knowledge. In *Proceedings of* *the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2139–2145, Online. Association for Computational Linguistics.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hitomi Yanaka and Koji Mineshima. 2022. Compositional evaluation on Japanese textual entailment and similarity. *Transactions of the Association for Computational Linguistics*, 10:1266–1284.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. ZeroGen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ann Yuan, Daphne Ippolito, Vitaly Nikolaev, Chris Callison-Burch, Andy Coenen, and Sebastian Gehrmann. 2021. Synthbio: A case study in faster curation of text datasets. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

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.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Hao Zhang, Youlin Wu, Junyu Lu, Zewen Bai, Jiangming Wu, Hongfei Lin, and Shaowu Zhang. 2023a. ZBL2W at SemEval-2023 task 9: A multilingual fine-tuning model with data augmentation for tweet intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 770–775, Toronto, Canada. Association for Computational Linguistics.

Hongming Zhang, Daniel Khashabi, Yangqiu Song, and Dan Roth. 2020a. Transomcs: From linguistic graphs to commonsense knowledge. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4004–4010. International Joint Conferences on Artificial Intelligence Organization. Main track.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020b. Aser: A large-scale eventuality knowledge graph. In *Proceedings of The Web Conference 2020*, WWW '20, page 201–211, New York, NY, USA. Association for Computing Machinery.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. Record: Bridging the gap between human and machine commonsense reading comprehension.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023b. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods* in Natural Language Processing, pages 7915–7927, Singapore. Association for Computational Linguistics.

Ziyin Zhang, Yikang Liu, Weifang Huang, Junyu Mao, Rui Wang, and Hai Hu. 2023c. Mela: Multilingual evaluation of linguistic acceptability.

A Details of the Experimental Settings

We used mBERT (Devlin et al., 2019), XLM-100 (Conneau and Lample, 2019), XLM-R (Conneau et al., 2020), and mDeBERTa-v3 (He et al., 2023) as encoder-based multilingual LMs, Llama2-70B (Touvron et al., 2023), GPT-3.5 (Ouyang et al., 2022), and GPT-4 (OpenAI et al., 2024) as decoder-based multilingual LMs for the experiments. Table 9 shows the details of the LMs. Encoder-based LMs were fine-tuned following the settings in Table 10. Decoder-based LMs inferred with 0-shot and 3-shot settings[12] with a fixed seed value. For GPT-3.5 and GPT-4, top_p and temperature were set to 0 to achieve as deterministic outputs as possible. For Llama2-70B, output was generated greedy, and outlines (Willard and Louf, 2023) were used to fix the output format.

Type	Model Name	HuggingFace / OpenAI API
Encoder	mBERT-cased	bert-base-multilingual-cased
	mBERT-uncased	bert-base-multilingual-uncased
	XLM-100	xlm-mlm-100-1280
	XLM-R$_{\text{BASE}}$	xlm-roberta-base
	XLM-R$_{\text{LARGE}}$	xlm-roberta-large
	mDeBERTa-v3	microsoft/mdeberta-v3-base
Decoder	Llama2-70B	meta-llama/Llama-2-70b-chat-hf
	GPT-3.5	gpt-3.5-turbo-1106
	GPT-4	gpt-4-1106-preview

Table 9: Details of the LMs for the experiments.

Hyper-parameter	Value
Batch Size	64
Learning Rate	2e-5, 3e-5, 5e-5
Seed	42
Early Stopping	3
Warmup Ratio	0.1
Max Sequence Length	128

Table 10: The hyper-parameters used in the experiment, and others, were set to default settings. The implementation used Transformers (Wolf et al., 2020).

B Qualitative Analysis of mCSQA

Table 11 shows examples of mCSQAs for each language. The examples in Table 11 are accompanied by English translations using DeepL[13] to avoid arbitrary translation.

B.1 Can Multilingual LMs Take into Account Language-specific Knowledge?

Case study When we examine some cases in Table 11, such as the examples from the Dutch Hard set and the Russian Hard set, we find that the English translations contain duplications among the question choices. However, these duplications arise not from differences in tense or conjugation, but from semantic differences unique to each language, which a native speaker, equipped with language-specific knowledge and common sense, could easily distinguish. Furthermore, in the case of the German Easy sets, knowledge of Germany's unique education system is required, which might be challenging for those unfamiliar with it. Yet, for German speakers, it is common knowledge that such education systems, such as the Abitur[14] related to the Gymnasium[15], making it answerable for those knowledgeable in German. This demonstrates that multilingual LMs are capable of generating questions that include the kind of language-specific knowledge and common sense that a native speaker would possess.

The effectiveness of the CSQA style QA When examining the Japanese Hard set in Table 10, all the choices translate into the names of seafood in English, which does not match the context of a female singer mentioned in the question. Japanese native speakers would normally recognize them as seafood names too, making it seem at first glance that there is no correct answer. However, the correct choice, 'あゆ' (ayu), when pronounced in Japanese, is read as 'ayu'. This pronunciation is widely known across Japan as the nickname for the famous singer '浜崎あゆみ' (Ayumi Hamasaki)[16], making it a plausible choice even though it's not strictly correct. It allows for a satisfactory selection by Japanese native speakers with language-specific knowledge, common sense, and cultural awareness, and is not answerable by English translation only. In Japan, nicknames are often derived from abbre-

[12] In the 3-shot setting, the examples were selected randomly from the training data and included both Easy and Hard sets.

[13] We are using DeepL Pro (https://www.deepl.com/translator), therefore, the copyright of the translations belongs to us. (https://www.deepl.com/pro-license)

[14] https://en.wikipedia.org/wiki/Abitur

[15] https://en.wikipedia.org/wiki/Gymnasium_(school)

[16] https://en.wikipedia.org/wiki/Ayumi_Hamasaki

viations of their names or can suggest the names of objects. The distractor ‘Wakame’ is known as the name of a character from the long-running, famous anime ‘Sazae-san’[17] but not as a singer, thus serving its purpose as a distractor in this question effectively. Similarly, if there were a choice like ‘いくら’ (common meaning: red caviar; pronounced: ikura), the plausibility of choice in this question might have been divided. Recently, ‘ikura’[18] has become a popular name, associated with a member of ‘Yoasobi’[19], a popular artist group among young people. Adding such a choice would confuse the choice of the correct answer because both choices are plausible, so it would not serve effectively as a distractor. This case shows that the choices can define the scope of common sense, thus making the question effective in evaluating common sense accurately.

B.2 The Relationship between Knowledge, Culture, Commonsense, and Social Bias

B.2.1 What is the Commonsense?

As can be seen from Table 11 and the discussions in section B.1, language-specific common sense is closely related to knowledge and culture. The ConceptNet used in this study does not limit the scope of common sense and deals with a wide range of common sense, enabling the inclusion of questions from various backgrounds into mCSQA, following the same trend as CSQA and JCSQA.

Generally, commonsense not based on the specific culture or knowledge of a language is likely to be a common understanding across all languages, making such problems potentially answerable through the language-transfer ability of multilingual LMs. However, as shown in Table 1, the granularity of actions, events, and behaviors differs by language, which can be considered to be influenced by the cultural background of the language area.

This study focuses on language-specific common sense that cannot be addressed by translations of datasets from other languages, and the culture and knowledge included in them are shared among native speakers. Therefore, answering questions that require language-specific backgrounds necessitates a certain level of knowledge and culture specific to each language. However, content that is too specialized falls outside the scope of common sense, and common sense and backgrounds vary among individuals. Therefore, we emphasize the precision of coverage in the manual question quality verification steps and employ a majority vote baseline to avoid overly relying on specific knowledge or culture.

In this way, questions were created that have language-specific common sense which is general for native speakers but not too specialized. If there was a need to create questions asking for knowledge specialized in specific fields, other knowledge bases such as ATOMIC (Sap et al., 2019a), and CCSK (Nguyen et al., 2023) could be used. However, this study focused on multilingual performance, deeming ConceptNet appropriate for mCSQA.

B.2.2 Is Commonsense Social Bias?

Since commonsense includes implicit cognition, it may contain social and cultural biases, and some methods for the removal of explicit and implicit social biases have been proposed (Sap et al., 2020; Field and Tsvetkov, 2020; Huang et al., 2021; Lent and Søgaard, 2021; Emelin et al., 2021; Bauer et al., 2023).

Social Chemistry 101 (Forbes et al., 2020), BBQ (Parrish et al., 2022), and SODAPOP (An et al., 2023) have been proposed for identifying biases within models or for bias detection using LMs. However, it remains challenging to address situations where biased thinking may only emerge when considering multiple-choice QA, where bias does not occur in isolation.

The definition of bias and common sense changes over time and varies from society to society, and what is considered common sense can shift to bias (Lee et al., 2023). Therefore, regular updates to the commonsense reasoning datasets are necessary. Our method for generating commonsense reasoning task datasets using LMs allows for low-cost update operations, making it possible to adapt to the changing boundaries between common sense and bias over time. However, this does not fundamentally address the inclusion of bias in datasets. Moreover, such issues require a deep chain of semantic thinking for resolution, making filtering based on textual information inappropriate. Therefore, it is necessary to develop methods to remove potential biases in commonsense reasoning task datasets in future work.

[17] https://en.wikipedia.org/wiki/Sazae-san
[18] https://en.wikipedia.org/wiki/Lilas_Ikuta
[19] https://en.wikipedia.org/wiki/Yoasobi

C Discoveries about the LMs Capabilities

C.1 Can LMs Create Questions including Commonsense?

Generation capability CommonGen (Lin et al., 2020) is one of the commonsense reasoning datasets that evaluates whether it is possible to create commonsense sentences from a given set of keywords. According to the leaderboard of CommonGen[20], the performance of GPT-3.5 used in our dataset creation demonstrates a capability for generating commonsense sentences comparable to those written by humans. However, there is still room for improvement in aspects such as word order. Therefore, we introduced refinement steps to encourage corrections in word order and other errors. Since language models have high performance in Grammar Error Correction (GEC) (Loem et al., 2023; Sottana et al., 2023; Fang et al., 2023; Coyne et al., 2023; Kaneko and Okazaki, 2023; Kwon et al., 2023), combining sentence generation from keywords with GEC capabilities in a pipeline helps to compensate for the weaknesses of language models. We believe that the quality of mCSQA questions is at least not inferior to those created by crowd-workers. The capability of multilingual LMs to create commonsense sentences from given keywords has also been demonstrated in the Korean CommonGen (Seo et al., 2022), indicating that it is possible to generate commonsense sentences multilingually.

Ensuring the quality of questions In this study, we have created commonsense reasoning dataset questions using keywords extracted from ConceptNet. Therefore, the language-specific knowledge and commonsense for each language are guaranteed by ConceptNet. Moreover, the LM creates questions following the given instructions through its emergent capabilities from each keyword. To enhance the language-specific performance of the multilingual LM for each language, we have created prompts for each language in this study. As can be seen from the discussion in section B.1 and Table 11, it has become possible to generate questions that possess language-specific knowledge. One of the reasons for the capability to create questions with language-specific knowledge may be attributed to the training data of the LM. For example, Wikipedia, one of the common training data for LMs, has each language which contains descriptions of knowledge unique to that language, so by posing questions in each language, it is thought that knowledge specific to each language is invoked, enabling the generation of questions based on the knowledge of each language. However, this is a hypothesis, and further analysis will be necessary for verification in future work. Moreover, we have added distractors in addition to the keywords used for generating the question, which means that even if a question can be generated, it may not necessarily be answerable. Furthermore, questions that cannot be answered have been removed, thus ensuring the difficulty and answerability of the QA.

C.2 Multilingual Capabilities

Is polyglot template effective? We translated the prompt to use question generation for each language and tuned it to convey the same meaning in each language in Section D aimed to emergence the language-specific knowledge. However, it is known that current generative LMs have mainly trained on English, which is better performance for queries made in English. However, several studies (Ahn et al., 2022; Shi et al., 2023; Wei et al., 2022b; Awasthi et al., 2023; Kasai et al., 2023; Jin et al., 2023) show enough performance even if multilingual queries. Note that the reported performance focuses on the ability to answer specific tasks on benchmarks and does not evaluate the emergent multilingual ability, especially question generation. Nevertheless, Whitehouse et al. (2023) shows that the text generation capability beyond English. As shown in Table 10, we were able to generate questions containing language-specific knowledge from the given keywords as intended by using prompts translated into each language. We were able to generate questions that require deep reasoning, including cultural backgrounds and language-specific pronunciation information as shown in Section B.1. Therefore, we conclude that using prompts tailored for each language is effective.

Is GPT-3.5 Multilingual LM? Yes, some studies (Lai et al., 2023; Armengol-Estapé et al., 2022; Zhang et al., 2023b) have indeed examined multilingual performance, and the training data also includes multilingually[21]. Therefore, the multilingual capabilities of GPT-3.5, GPT-4, and Llama used in our experiment have also been evalu-

[20] https://github.com/allenai/CommonGen-Eval

[21] https://github.com/openai/gpt-3/tree/master/dataset_statistics

ated (Ahuja et al., 2023; Schott et al., 2023; Chen et al., 2024), leading us to consider these as multilingual LMs. However, they still rely predominantly on information from Western norms (Cao et al., 2023; Arora et al., 2023; Havaldar et al., 2023), making this issue an ongoing challenge to be addressed in the future.

Exhortation to multilingual instruction-tuning dataset. Instruction-tuning (Wei et al., 2022a; Longpre et al., 2023; Chung et al., 2022; Wang et al., 2023) can enhance the quality of LMs, e.g. ability to follow instructions and NLU performance. However, in Section 2, the current multilingual datasets include those created through translation, which means that instruction-tuning using such data may not lead to the acquisition of data bias or language-specific knowledge. Given these considerations, the multilingual instruction-tuning data (Kew et al., 2023; Singh et al., 2024) proposed recently often utilize datasets created through translation, leading to the occurrence of the aforementioned issues to a considerable extent. Consequently, the effectiveness of such instruction-tuning may be diminished. For commonsense reasoning tasks in multilingual instruction-tuning datasets, they sometimes use X-CSQA (Lin et al., 2021). However, since it cannot handle language-specific knowledge or commonsense effectively, it is preferable to use data created from scratch, like mCSQA. Currently, due to data resource issues, reliance on translated data is inevitable, but we hope that in the future, it will be replaced by language-specific data.

C.3 Hard Sets are Truly Hard?

The Hard sets consist of questions that the LM used for question creation could not answer, thus reflecting the characteristics of that LM. However, despite the influence of specific LM's character, a performance decline in the Hard sets compared to the Easy sets was observed across all models. Therefore, while the strict division of sets depends on the model, it has become clear that there is a similar trend across LMs as a whole. For this reason, scoring is conducted without distinguishing between Easy and Hard, using a total score for the entire set, which allows for the absorption of differences due to the models.

C.4 Generation Bias and Annotation Artifacts

It has been pointed out that datasets created by LMs contain generation bias (Omura et al., 2020; Zellers et al., 2019; Tamborrino et al., 2020), and those created by crowd-workers include specific patterns (Annotation Artifacts) (Gururangan et al., 2018; Chen et al., 2019; Omura et al., 2020). Annotation artifacts, in particular, have been noted in natural language inference tasks such as MNLI (Williams et al., 2018) and SNLI (Bowman et al., 2015), where choices can be easily distinguished by superficial words like "not".

However, Tamborrino et al. (2020) show that the impact of Annotation Artifacts is not present in the CSQA task. Similarly, in this study, we have separated question generation ability and answering ability during the question generation process and shuffled the options, so there are no clues included in the dataset. Moreover, we create Hard sets, even if such biased questions existed, the evaluation is conducted without these biases, allowing for an evaluation that removes these biases.

D Prompts for Creating mCSQA

The prompts used for creating mCSQA are presented as follows: English in Table 12, Japanese in Table 13, Chinese in Table 14, German in Table 15, Portuguese in Table 16, Dutch in Table 17, French in Table 18 and Russian in Table 19.

In each prompt template, the words within the curly brackets are replaced with data-specific terms[22] before input to the LM.

Furthermore, as discussed in Section C.2, each template was translated exactly to elicit language-specific knowledge of each language. The translations were carried out using both GPT-3.5 and DeepL to ensure there were no semantic differences, with manual fixing applied as needed. We use the OpenAI API's JSON mode[23] has facilitated the retrieval of generation results.

Our findings as a tip, when inputting structured data such as keywords, doing so in a format similar to a programming code like list type, allows us to obtain results that more following the prompt instructions. This improvement can be attributed to the LM's learning to enhance coding abilities, which is believed to have improved its recognition capabilities.

[22]https://peps.python.org/pep-0498/

[23]https://platform.openai.com/docs/guides/text-generation/json-mode

Lang.		Question	Choices				
			Correct	Distractors		Additional Distractors	
EN	Easy	If a cat is feeling irritated, what might it do?	scratch if annoy	look out window	fish with paw	chase a toy	nap in the sun
	Hard	Which animal is known for its playful behavior and agile movements?	monkey	jellyfish	lemur	orangutan	gorilla
JA	Easy	音を聞き分けるためには何をしますか？ (What do you do to listen to the sounds?)	耳を澄ます (Listen carefully)	学習する (Learn)	書き取る (Write)	実践する (Practice)	経験する (Experience)
	Hard	日本の女性歌手で、自身の楽曲の作詞・作曲も手がける人気アーティストは誰ですか？ (Which popular Japanese female singer also writes lyrics and composes her own songs?)	あゆ (Sweetfish)	どじょう (Loach)	わかめ (Wakame)	うなぎ (Eel)	さけ (Salmon)
ZH	Easy	你在考试前应该做什么? (What should you do before your exam?)	回家温习 (Go home and study)	聊天 (Chatting)	作弊 (Cheat)	健身 (Work out)	看电影 (Watch films)
	Hard	在感情关系中，最令人痛苦的事情是什么? (What's the most excruciating thing about being in a relationship?)	被甩 (Getting dumped)	花大钱 (Spending a lot of money)	心碎 (Getting your heart broken)	找到真爱 (Finding true love)	实现梦想 (Realising your dreams)
DE	Easy	Welche Art von weiterführender Schule bereitet Schüler auf das Abitur vor? (What type of secondary school prepares students for the Abitur?)	gymnasium (grammar school)	gesamtschule (comprehensive school)	fachoberschule (technical seco-ndary school)	berufsschule (vocational school)	realschule (secondary school)
	Hard	Was ist die richtige Bezeichnung für das langsame Abwärtsbewegen auf einer schiefen Ebene? (What is the correct term for moving slowly downwards on an inclined plane?)	hinabgleiten (slide down)	hinabfliegen (fly down)	dahinab (descend)	hinabtauchen (dive down)	hinabschweben (float down)
PT	Easy	Como demonstrar afeto a um animal de estimação? (How do you show affection to a pet?)	fazer carinho (cuddle)	alegrar a vida (combing)	pentelhar (brighten up life)	abraçar (cuddle)	dar um presente (give a gift)
	Hard	Qual a ação que um coelho pode fazer para se mover rapidamente? (What action can a rabbit do to move quickly?)	pular (jump)	orientando (guiding)	segurar (hold)	esperar (wait)	correr (run)
NL	Easy	Kunt u mij vertellen wat gokken is? (Can you tell me what gambling is?)	kansspel (game of chance)	gelijkspel (draw)	steekspel (joust)	vuurspel (match)	wedstrijd (fire game)
	Hard	Kunt u uitleggen wat een veelvoorkomend begrip is dat verwijst naar iets wat algemeen geaccepteerd of verspreid is in een samenleving? (Can you explain what is a common term that refers to something commonly accepted or widespread in a society?)	gemeengoed (common)	gemeenschap (community)	gemeenplaats (commonplace)	gezamenlijk (common)	gebruikelijk (common)
FR	Easy	Quelle unité de temps correspond à une période de vingt-quatre heures ? (What unit of time corresponds to a twenty-four hour period?)	jour (day)	décade (decade)	siècle (century)	année (year)	mois (month)
	Hard	Quelle partie du corps utilise-t-on pour saisir des objets de petite taille ? (What part of the body is used to pick up small objects?)	doigt (finger)	annulaire (ring finger)	auriculaire (little finger)	majeur (middle finger)	index (index finger)
RU	Easy	Какое время года обычно связывается с праздниками Нового года и Рождества? (What time of year is usually associated with the holidays of New Year's Eve and Christmas?)	зима (winter)	весна (spring)	осень (fall)	летний сезон (summer season)	лето (summer)
	Hard	Какой звук издает довольный кот? (What sound does a contented cat make?)	урчание (purr)	заурчать (rumble)	проурчать (purr)	мурлыкать (purr)	громко урчать (purr)

Table 11: The examples of mCSQA. The English translations are all machine-translated by DeepL. The translated results sometimes are aggregated into one English word due to ignoring source language-specific subtle meaning differences caused by machine translation. This aggregation has also been observed in X-CSQA, which was created using machine translation of CSQA. Hence, X-CSQA could not evaluate fine-grained, language-specific knowledge for each language, but mCSQA can evaluate it because it is created from scratch for each language.

Steps	Prompt (English)
Create question sentences	Please create a multiple-choice question with the following conditions: (a) The only correct answer is ["{correct}"]. (b) The incorrect answers are ["{distractor1}", "{distractor2}"]. (c) Do not use the words ["{correct}", "{distractor1}", "{distractor2}"] in the question. (d) Avoid using superficial information, such as character count. (e) The question ends with a question mark (?). (f) It should be an objective question that can be sufficiently answered with common sense knowledge alone. (g) The question must be a simple and short sentence consisting of only one sentence. Question:
Refine question sentences	If the original sentence is semantically and grammatically correct, repeat it; if it is unnatural, please rewrite it into a correct and fluent sentence. {question}
Add additional distractors	Please only add two plausible and natural choices and save them in {'additional_choice':[]}. ["{choice1}", "{chioce2}", "{choice3}"]
Verify Qualities	Please select only one alphabet as the answer from the Answer Choices and save it in the format: {'answer': selected_answer}. Q: {question} Answer Choices: (A) {choice_a} (B) {choice_b} (C) {choice_c} (D) {choice_d} (E) {choice_e}

Table 12: The prompt templates used to create the mCSQA in the English version.

Steps	Prompt (Japanese)
Create question sentences	以下の条件を満たす選択肢付きのクイズ問題を作成してください。 (a) 正解は["{correct}"]のみです。 (b) 不正解は["{distractor1}", "{distractor2}"]です。 (c) 問題文に["{correct}", "{distractor1}", "{distractor2}"]という単語を使わないでください。 (d) 文字数などの表面的な情報の使用を避けてください。 (e) 問題文は疑問符（？）で終わります。 (f) 一般常識だけで十分に答えられる客観的な問題である必要があります。 (g) 問題文は一文のみから成る単純で短い文でなければなりません。 問題：
Refine question sentences	元の文が意味的・文法的に正しい場合は繰り返す、 不自然な場合は正しい流暢な文へ書き換えてください。 {question}
Add additional distractors	もっともらしい自然な選択肢を2つだけ追加し、 それらを{'additional_choice':[]}に保存してください。 ["{choice1}", "{chioce2}", "{choice3}"]
Verify Qualities	Answer Choicesから解答となるアルファベットを 1 つだけ選び、 次の形式で保存してください：{'answer': selected_answer}。 Q: {question} Answer Choices: (A) {choice_a} (B) {choice_b} (C) {choice_c} (D) {choice_d} (E) {choice_e}

Table 13: The prompt templates used to create the mCSQA in the Japanese version.

Steps	Prompt (Chinese)
Create question sentences	请根据以下条件创建一个多项选择题: (a) 唯一正确答案是["{correct}"]。 (b) 错误答案是["{distractor1}", "{distractor2}"]。 (c) 问题中不得使用["{correct}", "{distractor1}", "{distractor2}"]这些词。 (d) 避免使用表面信息，如字符数。 (e) 问题以问号(?)结束。 (f) 它应该是一个客观的问题，仅凭常识就能充分回答。 (g) 问题必须是一个简单且短的句子，仅由一句话组成。 问题:
Refine question sentences	如果原句在语义和语法上正确，请重复它；如果不自然，请将其改写为正确流畅的句子。 {question}
Add additional distractors	请只添加两个合理且自然的选择，并将它们保存在 {'additional_choice':[]} 中。 ["{choice1}", "{chioce2}", "{choice3}"]
Verify Qualities	请从Answer Choices中仅选择一个字母作为答案，并以以下格式保存: {'answer': selected_answer}。 Q: {question} Answer Choices: (A) {choice_a} (B) {choice_b} (C) {choice_c} (D) {choice_d} (E) {choice_e}

Table 14: The prompt templates used to create the mCSQA in the Chinese version.

Steps	Prompt (German)
Create question sentences	Bitte erstellen Sie eine Multiple-Choice-Frage mit folgenden Bedingungen: (a) Die einzig richtige Antwort ist ["{correct}"]. (b) Die falschen Antworten sind ["{distractor1}", "{distractor2}"]. (c) Verwenden Sie in der Frage nicht die Wörter ["{correct}", "{distractor1}", "{distractor2}"]. (d) Vermeiden Sie oberflächliche Informationen, wie z.B. die Zeichenanzahl. (e) Die Frage endet mit einem Fragezeichen (?). (f) Es sollte eine objektive Frage sein, die allein mit Allgemeinwissen ausreichend beantwortet werden kann. (g) Die Frage muss ein einfacher und kurzer Satz bestehend aus nur einem Satz sein. Frage:
Refine question sentences	Wenn der Originalsatz semantisch und grammatikalisch korrekt ist, wiederholen Sie ihn; wenn er unnatürlich ist, schreiben Sie ihn bitte in einen korrekten und flüssigen Satz um. {question}
Add additional distractors	Bitte fügen Sie nur zwei plausible und natürliche Optionen hinzu und speichern Sie diese in {'additional_choice':[]}. ["{choice1}", "{chioce2}", "{choice3}"]
Verify Qualities	Bitte wählen Sie nur einen Buchstaben als Antwort aus den Answer Choices aus und speichern Sie ihn im Format: {'answer': selected_answer}. Q: {question} Answer Choices: (A) {choice_a} (B) {choice_b} (C) {choice_c} (D) {choice_d} (E) {choice_e}

Table 15: The prompt templates used to create the mCSQA in the German version.

Steps	Prompt (Portuguese)
Create question sentences	Por favor, crie uma pergunta de múltipla escolha com as seguintes condições: (a) A única resposta correta é ["{correct}"]. (b) As respostas incorretas são ["{distractor1}", "{distractor2}"]. (c) Não use as palavras ["{correct}", "{distractor1}", "{distractor2}"] na pergunta. (d) Evite usar informações superficiais, como a contagem de caracteres. (e) A pergunta termina com um ponto de interrogação (?). (f) Deve ser uma pergunta objetiva que pode ser suficientemente respondida apenas com conhecimento de senso comum. (g) A pergunta deve ser uma frase simples e curta, consistindo de apenas uma frase. Pergunta:
Refine question sentences	Se a frase original estiver semanticamente e gramaticalmente correta, repita-a; se for pouco natural, por favor, reescreva-a em uma frase correta e fluente. {question}
Add additional distractors	Por favor, adicione apenas duas escolhas plausíveis e naturais e salve-as em {'additional_choice':[]}. ["{choice1}", "{chioce2}", "{choice3}"]
Verify Qualities	Por favor, selecione apenas uma letra como resposta das Answer Choices e salve no formato: {'answer': selected_answer}. Q: {question} Answer Choices: (A) {choice_a} (B) {choice_b} (C) {choice_c} (D) {choice_d} (E) {choice_e}

Table 16: The prompt templates used to create the mCSQA in the Portuguese version.

Steps	Prompt (Dutch)
Create question sentences	Maak alstublieft een meerkeuzevraag met de volgende voorwaarden: (a) Het enige juiste antwoord is ["{correct}"]. (b) De onjuiste antwoorden zijn ["{distractor1}", "{distractor2}"]. (c) Gebruik de woorden ["{correct}", "{distractor1}", "{distractor2}"] niet in de vraag. (d) Vermijd het gebruik van oppervlakkige informatie, zoals het aantal tekens. (e) De vraag eindigt met een vraagteken (?). (f) Het moet een objectieve vraag zijn die alleen met algemene kennis voldoende beantwoord kan worden. (g) De vraag moet een eenvoudige en korte zin zijn die uit slechts één zin bestaat. Vraag:
Refine question sentences	Als de originele zin semantisch en grammaticaal correct is, herhaal deze dan; als het onnatuurlijk is, herschrijf het dan naar een correcte en vloeiende zin. {question}
Add additional distractors	Voeg alstublieft slechts twee aannemelijke en natuurlijke keuzes toe en sla ze op in {'additional_choice':[]}. ["{choice1}", "{chioce2}", "{choice3}"]
Verify Qualities	Selecteer alstublieft slechts één letter als antwoord uit de Answer Choices en sla het op in het formaat: {'answer': selected_answer}. Q: {question} Answer Choices: (A) {choice_a} (B) {choice_b} (C) {choice_c} (D) {choice_d} (E) {choice_e}

Table 17: The prompt templates used to create the mCSQA in the Dutch version.

Steps	Prompt (French)
Create question sentences	Veuillez créer une question à choix multiples avec les conditions suivantes : (a) La seule bonne réponse est ["{correct}"]. (b) Les réponses incorrectes sont ["{distractor1}", "{distractor2}"]. (c) Ne pas utiliser les mots ["{correct}", "{distractor1}", "{distractor2}"] dans la question. (d) Évitez d'utiliser des informations superficielles, telles que le nombre de caractères. (e) La question se termine par un point d'interrogation (?). (f) Il doit s'agir d'une question objective qui peut être suffisamment répondue avec le seul sens commun. (g) La question doit être une phrase simple et courte composée d'une seule phrase. Question :
Refine question sentences	Si la phrase originale est correcte sémantiquement et grammaticalement, répétez-la ; si elle est peu naturelle, veuillez la reformuler en une phrase correcte et fluide. {question}
Add additional distractors	Veuillez ajouter seulement deux choix plausibles et naturels et les enregistrer dans {'additional_choice':[]}. ["{choice1}", "{chioce2}", "{choice3}"]
Verify Qualities	Veuillez sélectionner uniquement une lettre comme réponse parmi les Answer Choices et enregistrez-la dans le format : {'answer': selected_answer}. Q: {question} Answer Choices: (A) {choice_a} (B) {choice_b} (C) {choice_c} (D) {choice_d} (E) {choice_e}

Table 18: The prompt templates used to create the mCSQA in the French version.

Steps	Prompt (Russian)
Create question sentences	Пожалуйста, создайте вопрос с несколькими вариантами ответа с учетом следующих условий: (a) Единственный правильный ответ - ["{correct}"]. (b) Неправильные ответы - ["{distractor1}", "{distractor2}"]. (c) Не используйте слова ["{correct}", "{distractor1}", "{distractor2}"] в вопросе. (d) Избегайте использования поверхностной информации, такой как количество символов. (e) Вопрос заканчивается вопросительным знаком (?). (f) Это должен быть объективный вопрос, на который можно достаточно ответить только с помощью здравого смысла. (g) Вопрос должен быть простым и коротким, состоящим только из одного предложения. Вопрос:
Refine question sentences	Если исходное предложение семантически и грамматически правильно, повторите его; если оно звучит ненатурально, пожалуйста, перепишите его на корректный и свободно звучащий язык. {question}
Add additional distractors	Пожалуйста, добавьте только два правдоподобных и естественных выбора и сохраните их в {'additional_choice':[]}. ["{choice1}", "{chioce2}", "{choice3}"]
Verify Qualities	Пожалуйста, выберите только одну букву алфавита в качестве ответа из Answer Choices и сохраните её в формате: {'answer': selected_answer}. Q: {question} Answer Choices: (A) {choice_a} (B) {choice_b} (C) {choice_c} (D) {choice_d} (E) {choice_e}

Table 19: The prompt templates used to create the mCSQA in the Russian version.