mSCAN: A Dataset for Multilingual Compositional Generalisation Evaluation

Amélie Reymond
University of Washington
attr@uw.edu

Shane Steinert-Threlkeld
University of Washington
shanest@uw.edu

Abstract

Language models achieve remarkable results on a variety of tasks, yet still struggle with compositional generalisation benchmarks. The majority of these benchmarks evaluate performance in English only, leaving us with the question of whether these results generalise to other languages. As an initial step to answering this question, we introduce mSCAN, a multilingual adaptation of the SCAN dataset. It was produced by a rule-based translation, developed in cooperation with native speakers. We then showcase this novel dataset on some in-context learning experiments, with the multilingual large language model `BLOOM` as well as `gpt3.5-turbo`.

1 Introduction

Humans learn quickly by easily recombining previously known concepts in unseen settings. Several benchmarks have been designed to empirically investigate whether neural networks are equipped with similar abilities (Lake and Baroni, 2018; Keysers et al., 2020; Hupkes et al., 2020; Kim and Linzen, 2020). Such benchmarks are composed of tasks in which the training data and the test data have different and carefully chosen distributions. Recent work used these benchmarks to evaluate pretrained large language models (LLMs) and showed that despite their remarkable success on many other tasks they still struggle with compositional generalisation (Qiu et al., 2022).

The majority of the research on compositional generalisation has focussed on English data and models — but do compositional generalisation abilities differ across languages? Indeed, it has been argued that the performance of a model in English is not a guarantee that it will work "equally or even reasonably well" in other languages (Bender, 2011). On top of that, compositional generalisation itself is not guaranteed to work uniformly across human languages (Bittner, 1995).

Furthermore, the exploration of cross- and multilingual compositional generalisation could benefit the expansion of language technology to low-resource languages and settings (Chaabouni et al., 2021), as a potential approach to overcome the need for huge amounts of data that neural models require.

With ever-increased scale, some large language models have shown great performance on downstream tasks while only conditioned on a few examples, and without updating their parameters. This is known as in-context learning, a paradigm in which some very large models such as GPT-3 and PaLM have been shown to manifest reasoning abilities when prompted in specific ways, including in multilingual settings (Shi et al., 2022). Despite these promising perspectives, it does not currently stand as an alternative to fine-tuning. Some recent research has sought to investigate compositional generalisation within the in-context learning paradigm, showing it gets outperformed by smaller fine-tuned models.

As a means to further the study of compositional generalisation in multiple languages, we introduce mSCAN (multilingual SCAN), an adaptation of the SCAN benchmark into French, Hindi, Mandarin Chinese and Russian. We also provide for each language both the original SCAN benchmark splits (`add_jump, add_turn_left, length`) as well as the Maximum Compound Divergence splits (Keysers et al., 2020).

We also present preliminary experimental results using mSCAN in an in-context learning paradigm on `BLOOM` and `gpt3.5-turbo`.

Following the GenBench taxonomy (Hupkes et al., 2023), the primary motivation for this work can be characterised as intrinsic given its primary function to provide a means to evaluate compositional generalisation in multilingual settings. Similarly to the original SCAN benchmark, the source of the distribution shift is fully generated and its

Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP, pages 143–151

Motivation

Practical	*Cognitive*	*Intrinsic*	*Fairness*
		☐	

Generalisation type

Compositional	*Structural*	*Cross Task*	*Cross Language*	*Cross Domain*	*Robustness*
☐					

Shift type

Covariate	*Label*	*Full*	*Assumed*
☐			

Shift source

Naturally occuring	*Partitioned natural*	*Generated shift*	*Fully generated*
			☐

Shift locus

Train–test	*Finetune train–test*	*Pretrain–train*	*Pretrain–test*
			☐

Figure 1: GenBench evaluation card

type is covariate. Moreover, the in-context set-up of our experiments places the shift locus between the pre-train and test stages though we note that the data can also be used in a fine-tuning setup in the future.

2 Background

Pre-trained multilingual models seek to address the challenge of low-resource languages, by leveraging the pre-training and the hope that high-resource languages will help lower-resource ones. Large-scale multilingual language models have achieved impressive performance across typologically distinct languages (Ruder et al., 2021). Yet, the cross and within-language performance of downstream tasks on such models remain correlated to their amount of language-specific pertaining data (Lauscher et al., 2020).

However, if scaling up the amount of pre-training data might improve cross-lingual generalisation, it might come at a price when it comes to compositional generalisation. Kim et al. (2022) have questioned the reported benefits of pre-training on compositional generalisation benchmarks and have observed a case of inverse scaling, where the performance degradation on COGS actually increases with the amount of pre-training data.

In a further study on the impact of model scale on compositional generalisation, Qiu et al. (2022) compared fine-tuning, prompt-tuning and in-context learning on multiple compositional generalisation datasets and observed that for in-context learning, the performance is correlated with model size. However, it is worse than for fine-tuned, smaller models. Datasets they used included COGS and the Compositional Freebase Question dataset or CFQ (Keysers et al., 2020), which consists of questions and answers in natural language, as well as accompanying SPARQL queries against a knowledge base. (Qiu et al., 2022)

Hosseini et al. (2022) evaluated four model families for in-context learning on multiple semantic parsing benchmarks. Despite their observation that the larger models tend to do better, they report that the in-context learning performance on SCAN and CFQ is very small for the models tested.

MCWQ (Cui et al., 2022), a multilingual variant of CFQ, is the first adaptation into multiple languages of a compositional generalisation benchmark. It was created with the use of neural machine translation. Wang and Hershocovich (2023) have shown that using neural machine translation to translate already existing benchmarks entails "critical semantic distortion", and favour a rule-based translation of the MCWQ dataset.

The MSGM benchmark (Shi et al., 2022) investigates the mathematical reasoning abilities of LLMs in multilingual settings, by providing data in ten different languages. Even though the decomposition of SCAN commands closely resembles that of arithmetic operations, the MSGM differs in that it does not specifically target the capacity of the model to map forms to a representation of meaning. As such, there has not yet been any investigation specifically targeting the compositional generalisation abilities of multilingual models in an in-context setting.

3 The mSCAN dataset

Our goal was to adapt the Simplified version of the CommAi Navigation dataset or SCAN (Lake and Baroni, 2018) to languages that belong to typologically diverse families and typically are represented in varying proportions in the training data of multilingual models. The languages selected also have different language scripts: Latin, Cyrillic and Devanagari. The original SCAN consists of a set of navigation commands in English such as "jump left", and their corresponding sequence of actions, such as LTURN JUMP. It is a synthetic dataset: the natural language commands are generated by a phrase-structure grammar, and the actions are produced by applying a semantic interpretation function. As such, it is akin to a semantic parsing task.

3.1 Generation methodology: a grammar based-transduction

Following (Wang and Hershcovich, 2023), we translate SCAN in a rule-based manner.

The method we used consists of a set of English grammar rules, their accompanying transduction rules and word mappings.

We used the context-free grammar shown in Figure 2, which is exactly equivalent to the one from (Lake and Baroni, 2018), only differing in notation. We also used the interpretation function as provided in their work. The SCAN grammar does not have recursion and generates an unambiguous and finite set of 20910 natural language commands to action sequence pairs.

Native speakers of French, Mandarin Chinese, Russian and Hindi were asked to provide the corresponding interpretation function in their language. We consequently manually built the transduction functions, which were applied to the English parse trees. The resulting parse trees were then formed into our translated commands by word mappings.

For instance, for French translations, we first parsed the English text using the original SCAN grammar, given in Figure 2, to produce an English parse tree. This parse tree can be transduced into a French parse tree using the transduction rules given in Figure 3. These transduction rules tell us that, for instance, `S AND S` and `S AFTER S` should be translated word-for-word, and the translation of "and" is "et", and "after" is "après". They also tell us that French distinguishes between "turn left" (translated as "tourner à gauche") and "turn around left" (translated as "tourner autour par la gauche").

```
C -> S AND S | S AFTER S | S
S -> V TIMES | V
V -> ACTION VECTOR DIR
   | TURN VECTOR DIR
   | D | ACTION
D -> ACTION DIR | TURN DIR

ACTION -> 'walk' | 'look'
        | 'run' | 'jump'
TURN -> 'turn'
VECTOR -> 'around' | 'opposite'
DIR -> 'left' | 'right'
TIMES -> 'twice' | 'thrice'
AFTER -> 'after'
AND -> 'and'
```

Figure 2: English SCAN grammar

Upon the completion of generation, a sample was manually checked by the native speakers for meaning preservation.

3.2 Splits

We do not introduce a novel way to split our dataset and rather choose to directly reproduce already existing splits on mSCAN.

3.2.1 SCAN splits

The original SCAN dataset contains multiple types of splits, each aimed to test distinct levels of compositional ability: the "simple" split is a random subset of the data, and the "length" one targets commands with corresponding action sequences that are longer than any example seen during training, and finally, the "primitive" split, which tests whether a primitive only encountered in isolation during training can be used adequately novel combinations at test time.

3.2.2 Maximum Compound Divergence Splits

The MCD splits were introduced by (Keysers et al., 2020) with their distribution-based compositionality assessment (DBCA). It consists of a method to measure whether a dataset has been split adequately to test for compositional generalisation, as well as a method to construct such splits. The main principles of the DBCA are that (1) all the atoms or primitive elements existing in the test set should also be present in the training set, and in a distribution as similar as possible, and (2) that

```
# Non-terminals
[S AND S] -> [S] [AND] [S]
[S AFTER S] -> [S] [AFTER] [S]
...
[ACTION VECTOR DIR] -> [ACTION] [VECTOR]
     [DIR]
[ACTION LEFT] -> [ACTION] 'a gauche'
[ACTION RIGHT] -> [ACTION] 'a droite'
...
```

```
# Terminals
'and'   -> 'et'
'after' -> 'apres'
'turn'  -> 'tourner'
'right' -> 'par la droite'
'left'  -> 'par la gauche'
...
```

Figure 3: English to French transduction rules

the distribution of compounds (ways of composing the atoms) should be as different as possible between the training and the test set. Intuitively, this method seeks to ensure that what is measured is how the atoms are composed into new compounds and that the compositions are challenging enough so that the model cannot rely on anything else than its capacity to generalise compositionally.

(Keysers et al., 2020) applied MCD to SCAN, and we replicate these splits exactly in mSCAN: each line of the respective test, train and evaluation sets in mSCAN is a direct translation of the corresponding line in the English-language MCD SCAN split.

We make `mSCAN_fra`, `mSCAN_hin`, `mSCAN_rus` and `mSCAN_cmn` and their accompanying splits, available as a public dataset available on the Hugging Face platform[1].

4 Experiment

4.1 Models

The BigScience Large Open-Science Open-access Multilingual Language Model or `BLOOM`, (Workshop et al., 2023) is a Transformer-based language model with 176 billion parameters. As an autoregressive LLM, it is trained to generate text from a prompt. It was trained on 46 languages and 13 programming languages.

We also ran a small experiment on the OpenAI model `gpt3.5-turbo`, accessed via the OpenAI REST API, between 2023/10/23 and 2023/10/26.

4.2 Prompt design

Our approach focussed on the selection methodology of the in-context examples. Our goal was to adapt and mimic the principle underlying the original SCAN benchmark. That is, to test for compositional generalisation, the context examples should not contain the combinations of the test case.

[1]https://huggingface.co/datasets/CLMBR/mSCAN

We therefore randomly select the in-context examples from the training sets of our splits and the test case from the corresponding test sets. For example, a certain number of examples is sampled from the French `add_jump` training set, and its corresponding test case comes from the French `add_jump` test set. This example is cut out to only include the natural language commands and the start of the output sequence token ("OUT:"), therefore prompting the model to generate the adequate sequence of instructions as the output.

An EOS token was added at the end of each example and provided to the model as a stopping criterion parameter.

An example of a prompt is provided in Figure 4.

```
<s>IN: jump right thrice and turn
    opposite left OUT: I_TURN_RIGHT
    I_JUMP I_TURN_RIGHT I_JUMP
    I_TURN_RIGHT I_JUMP I_TURN_LEFT
    I_TURN_LEFT</s

<s>IN: walk after walk opposite left OUT
    : I_TURN_LEFT I_TURN_LEFT I_WALK
    I_WALK</s>

...

<s>IN: turn around left twice and look
    around left thrice OUT:
```

Figure 4: Example of a prompt in English

4.3 Set-up

Due to the context-size restrictions of the `BLOOM` model, we set the number of context examples to 8. In the original `add_primitive` SCAN splits, the primitive is over-represented in the training set by 10%. We imitate this in our set-up by manually adding the primitive to the context examples once, and by having removed the primitive from our train set, which ensures that the sampled remaining 7 in-context examples do not contain it. Therefore, the full prompt consists of 8 examples, of which one contains the primitive and 7 do not, and a test case that includes the primitive. We use greedy decoding for generation to provide a baseline.

5 Results

5.1 BLOOM

Because BLOOM was not trained with an end-of-sequence token, we truncated generated outputs to their expected length. Despite this adjustment, our results get zero exact match accuracies, that is, none of the full output sequences was equal to the correct answer. This is consistent with the results observed by Hosseini et al. (2022).

For a finer-grained measure of model performance than exact match accuracy, we measured the minimum edit distance between the truncated outputs and the target strings.

Table 1 shows the average minimum edit distance compared to the expected output length on 100 runs on the simple, MCD1, length and add_jump splits for each language. There is no result for `add_jump` on Hindi and Russian due to the encoding being larger than the maximum supported size for these experiments.

It is important to emphasise that there was no exact match, both for the original version of SCAN as well as for our mSCAN multilingual variants, meaning that the model has a 0% accuracy. We can observe however that there is a similar amount of error across languages.

As expected, the simple split achieves the best results, and Russian did not achieve a similar performance as the other languages, which are officially part of the BLOOM training corpus. Surprisingly, there is little difference between English and Hindi, while the model seems to do slightly better on Mandarin and French.

Despite Russian not being an official language part of the training data of BLOOM, we ran the experiments on our `mSCAN_rus` and we included it with the others.

5.2 GPT 3.5

Unlike with BLOOM, we obtained a few exact sequence matches with `gpt-3.5-turbo` but they are few, with less than 10% per language over the five languages including English. In this experiment again, Mandarin seems to achieve slightly better results. From these observations, it also appears that the model has the most difficulty with the `length` split.

The average edit distance results are better than those with BLOOM but display a similar pattern, with the model seeming to struggle the most on the `length` split and Mandarin achieving slightly better results. As expected, the model seems to be more successful with Russian than BLOOM.

6 Discussion

6.1 Pre-training data contamination

In the in-context set-up, the data from the pre-training corpus cannot be controlled. This means that there is a possibility that the compositional generalisation training set or the whole dataset itself could have been used. Given that BLOOM specifies the content of its training corpus, we are at least guaranteed that it has not learned the English SCAN dataset or that there was some test contamination. As we introduce mSCAN with this paper, it could not have been a part of the training data.

However, there is no guarantee the original SCAN has not been seen during the pre-training of the ChatGPT model. Given that we are not able to check the pre-training data, the data distribution shift is only *assumed* in this case.

6.2 In context-examples selection

It is acknowledged that prompting variations such as the format or order of prompts can have an influence on the in-context learning performance. Our context example selection methodology is rudimentary. A recent study found that the selection of in-context examples affects compositional generalisation performance, by showing that randomly selecting in-context leads to an accuracy gap compared to fine-tuned models (An et al., 2023). They argue that a careful selection of the in-context examples will "fully reveal the potential of in-context learning". They define three requirements for in-context examples: structural similarity, diversity and complexity. They show that this helps compositional generalisation. In the case of SCAN, the structural similarity factor is not as relevant, given the basic nature of the grammar (there are no complex structures such as in COGS). The diversity and complexity factors are not controlled in our experiment, given that we sample from the train set without looking at the number of distinct primitives included. For this reason, our set-up does not follow the principle that the primitives in the test

Model, language \ split		simple (13.55)	mcd1 (18.03)	length (30.04)	add_jump (14.58)
BLOOM	cmn	5.04	8.28	13.82	7.16
	eng	9.32	11.65	19.15	10.53
	fra	7.69	11.85	16.26	7.95
	hin	8.63	11.10	18.72	
	rus	12.04	15.60	27.21	
gpt-3.5-turbo	cmn	4.52	7.95	14.83	5.81
	eng	5.51	8.75	16.32	6.65
	fra	5.63	9.39	17.00	7.26
	hin	6.47	10.17	17.50	8.17
	rus	5.67	9.51	17.70	7.26

Table 1: Average edit distance for each language and split, on BLOOM and gpt-3.5-turbo. The numbers reported in the column headings correspond to the average expected output length. Note that BLOOM produced 0 exact matches.

Language \ split	simple	mcd1	length	add_jump
cmn	10	6	0	6
eng	7	7	0	1
fra	4	4	0	1
hin	0	0	1	2
rus	3	0	0	4

Table 2: Number of exact matches over 100 queries of gpt-3.5-turbo

case should be covered by the in-context examples. Instead, we expect the model to be able to infer the mapping to SCAN instructions from context as the instructions closely match their natural language counterparts (e.g., walk is mapped to I_WALK).

Other research uses a least-to-most prompting strategy: prompts consist of instructions telling explicitly the model to decompose the task into subproblems and showing it how to solve them sequentially (Zhou et al., 2023). The number of in-context examples in our experiment was constrained by the context size of the model in the BLOOM experiment. To work around this, the least-to-most method uses intermediate representations in the form of Python expressions, mapping for example "look twice" to "LOOK*2" instead of "LOOK LOOK". The authors show that the model is able to expand from the Python expression with high accuracy, but further investigation of the potential consequences of these intermediate representations could be pursued.

6.3 Compositional Generalisation and different languages

We observed that there was no large variation between how the different languages performed in our in-context setup, except for Mandarin Chinese, which has slightly better results. Given the limited scope of our experiments, this observation should be confirmed by further investigation. If these results hold then, they would be in contrast with previous findings, where in some NLP tasks, generative models (including BLOOM) perform better on higher-resource languages and languages that are in the Latin script (Ahuja et al., 2023).

6.4 Possibilities for future work

In addition to investigating different strategies for in-context example selection and systematically conducting the experiments on a larger scale than what this work presents, future work could involve adapting more realistic natural language tasks to multiple languages. Indeed, the subset of natural language covered by SCAN is small and its interpretation is more akin to arithmetic expressions than naturally occurring language. As such, it does not make it possible to evaluate for more sophisticated linguistic abstraction (Kim and Linzen, 2020). Adapting COGS to other languages would be an extensive process, requiring the construction of language-specific grammars.

It would also be worth doing experiments with fine-tuning on multilingual models such as mBART (Liu et al., 2020) or mT5 (Xue et al., 2021).

A systematic study of the interactions between (a) the size of language-specific pretraining data, and (b) both compositional and cross-lingual generalisation, would be an important contribution.

7 Conclusion

The majority of the research on compositional generalisation is focussed on English, leaving open the question as to whether its findings can generalise across languages. As an initial step towards this exploration, we introduce mSCAN, a multilingual adaptation of the SCAN dataset, produced using rule-based translation, with rules developed in cooperation with native speakers. We then showcase this novel dataset on some in-context learning experiments, with the multilingual large language model BLOOM.

Limitations

Due to the synthetic nature of the SCAN dataset, the translations in other languages do not aim to capture naturalness or fluency.

This dataset was created with the aim of expanding compositional generalisation evaluation to multiple languages. We evaluate BLOOM, a model carefully designed for multilingualism, trained on a meticulously curated corpus. Despite these two points, more typologically diverse and low-resource languages are absent from our dataset and our evaluation.

Finally, the scale of the experiments reported in this paper was limited by different factors, including the cost and time of inference, and the maximum context size of 1000 tokens of BLOOM. As such, larger-scale experiments would be needed to form a basis for comparison with other benchmark results.

Acknowledgements

We would like to thank the native speakers for their precious help with this work: Sofia Ahmed, Catalin Gangalic, Qingxia Guo, Sanjana Sharma, Onur Tuna, and Shengqi Zhu.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual Evaluation of Generative AI. ArXiv:2303.12528 [cs].

Shengnan An, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Jian-Guang Lou, and Dongmei Zhang. 2023. How Do In-Context Examples Affect Compositional Generalization? ArXiv:2305.04835 [cs].

Emily M. Bender. 2011. On Achieving and Evaluating Language-Independence in NLP. *Linguistic Issues in Language Technology*, 6.

Maria Bittner. 1995. Quantification in eskimo: A challenge for compositional semantics. In E. Bach, E. Jelinek, A. Kratzer, and B. Partee, editors, *Quantification in Natural Languages*, pages 59–80. Kluwer Academic Publishers.

Rahma Chaabouni, Roberto Dessì, and Eugene Kharitonov. 2021. Can transformers jump around right in natural language? assessing performance transfer from SCAN. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 136–148, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ruixiang Cui, Rahul Aralikatte, Heather Lent, and Daniel Hershcovich. 2022. Compositional Generalization in Multilingual Semantic Parsing over Wikidata. ArXiv:2108.03509 [cs].

Arian Hosseini, Ankit Vani, Dzmitry Bahdanau, Alessandro Sordoni, and Aaron Courville. 2022. On the Compositional Generalization Gap of In-Context Learning. ArXiv:2211.08473 [cs].

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality Decomposed: How do Neural Networks Generalise? *Journal of Artificial Intelligence Research*, 67:757–795.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. 2023. A taxonomy and review of generalization research in NLP. *Nature Machine Intelligence*, 5(10):1161–1174.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. 2020. Measuring Compositional Generalization: A Comprehensive Method on Realistic Data. ArXiv:1912.09713 [cs, stat].

Najoung Kim and Tal Linzen. 2020. COGS: A Compositional Generalization Challenge Based on Semantic Interpretation. ArXiv:2010.05465 [cs].

Najoung Kim, Tal Linzen, and Paul Smolensky. 2022. Uncontrolled Lexical Exposure Leads to Overestimation of Compositional Generalization in Pretrained Models. ArXiv:2212.10769 [cs].

Brenden Lake and Marco Baroni. 2018. Generalization without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2873–2882. PMLR. ISSN: 2640-3498.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

Linlu Qiu, Peter Shaw, Panupong Pasupat, Tianze Shi, Jonathan Herzig, Emily Pitler, Fei Sha, and Kristina Toutanova. 2022. Evaluating the Impact of Model Scale for Compositional Generalization in Semantic Parsing. ArXiv:2205.12253 [cs].

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards More Challenging and Nuanced Multilingual Evaluation. ArXiv:2104.07412 [cs].

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. 2022. Language models are multilingual chain-of-thought reasoners.

Zi Wang and Daniel Hershcovich. 2023. On Evaluating Multilingual Compositional Generalization with Translated Datasets. ArXiv:2306.11420 [cs].

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer.

Denny Zhou, Nathanael Scharli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. LEAST-TO-MOST PROMPTING ENABLES COMPLEX REASONING IN LARGE LANGUAGE MODELS.