It's All in the Heads: Using Attention Heads as a Baseline for Cross-Lingual Transfer in Commonsense Reasoning

Alexey Tikhonov*
Yandex
Berlin, Germany
altsoph@gmail.com

Max Ryabinin*
Yandex, HSE University
Moscow, Russia
mryabinin0@gmail.com

Abstract

Commonsense reasoning is one of the key problems in natural language processing, but the relative scarcity of labeled data holds back the progress for languages other than English. Pretrained cross-lingual models are a source of powerful language-agnostic representations, yet their inherent reasoning capabilities are still actively studied. In this work, we design a simple approach to commonsense reasoning which trains a linear classifier with weights of multi-head attention as features. To evaluate this approach, we create a multilingual Winograd Schema corpus by processing several datasets from prior work within a standardized pipeline and measure cross-lingual generalization ability in terms of out-of-sample performance. The method performs competitively with recent supervised and unsupervised approaches for commonsense reasoning, even when applied to other languages in a zero-shot manner. Also, we demonstrate that most of the performance is given by the same small subset of attention heads for all studied languages, which provides evidence of universal reasoning capabilities in multilingual encoders.

1 Introduction

Neural networks have achieved remarkable progress in numerous tasks involving natural language, such as machine translation (Bahdanau et al., 2014; Kaplan et al., 2020; Arivazhagan et al., 2019), language modeling (Brown et al., 2020), open-domain dialog systems (Adiwardana et al., 2020; Roller et al., 2020), and general-purpose language understanding (Devlin et al., 2019; He et al., 2021). However, the fundamental problem of commonsense reasoning has proven to be quite challenging for modern methods and arguably remains unsolved up to this day. The tasks that aim to measure reasoning capabilities, such as the Winograd Schema Challenge (Levesque et al., 2012), are deliberately designed not to be easily solved by statistical approaches, which are a foundation of most deep learning methods. Instead, these tasks require implicit knowledge about properties of real-world entities and their relations in order to resolve inherent ambiguities of natural language.

The town councilors refused to give *the demonstrators* a permit because they feared violence.
Answer: The town councilors

Figure 1: Example of a Winograd Schema problem. The resolved pronoun is underlined, two options are highlighted with an italic font.

Figure 1 illustrates the gist of this task: given a sentence and a pronoun (they), the goal is to choose the word that this pronoun refers to from two options (*The town councilors* or *the demonstrators*). While picking the right answer is straightforward for humans, the lack of explicit clues makes it hard for machine learning algorithms to perform better than majority vote or random choice.

Recently large Transformer-based masked language models (MLMs) (Devlin et al., 2019) were shown to achieve impressive results on several benchmark datasets for commonsense reasoning (Sakaguchi et al., 2020; Kocijan et al., 2019; Klein and Nabi, 2020). However, the best-performing methods frequently involve finetuning the entire model on large enough corpora with varying degrees of supervision; apart from providing initial parameter values, the pretrained trained language model is not used for predictions.

Moreover, these methods have mostly been evaluated on English language datasets, despite increasing interest in multilingual evaluation for NLP (Hu et al., 2020) and the existence of multilingual encoders (Conneau et al., 2020; Conneau and Lample, 2019). The XCOPA dataset (Ponti et al., 2020) was recently proposed as a benchmark for multilingual

*Equal contribution.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3534–3546
August 1–6, 2021. ©2021 Association for Computational Linguistics

commonsense reasoning, yet its task is different from the pronoun resolution problem described above. Versions of Winograd Schema Challenge exist in different languages, but each version comes with slight differences in task specification. This makes holistic cross-lingual evaluation of new commonsense reasoning approaches a quite difficult problem for researchers in the area.

In this work, we propose a simple supervised method for commonsense reasoning, which trains a linear classifier on the self-attention weights between the pronoun and two answer options. To evaluate our method and facilitate research in multilingual commonsense reasoning, we aggregate existing Winograd Schema datasets in English, French, Japanese, Russian, Portuguese, and Chinese languages, converting them to a single format with a strict task definition. Our approach performs comparably to supervised and unsupervised baselines in this setting with both multilingual BERT and XLM-R models as backbone encoders.

Moreover, we find that the same set of attention heads can be used to solve reasoning tasks in all languages, which hints at the emergence of language-independent linguistic functions in cross-lingual models and supports the conclusions made by prior work (Chi et al., 2020; Li et al., 2020). Interestingly, when using an unsupervised attention-based method (Klein and Nabi, 2019), we observe that restricting the choice of heads to this set also improves the results of this baseline. This result suggests that the key to improved performance of such approaches might lie in the right choice of heads rather then the exact attention values.

To summarize, our contributions are as follows:

- We offer a simple supervised method to utilize self-attention heads of pretrained language models for commonsense reasoning.
- We compile a multilingual dataset of Winograd schemas in six languages, bringing all tasks to the same format[1]. When evaluated on this dataset, our method performs competitively to strong baselines from prior work.
- We demonstrate that in cross-lingual models, there exists a small subset of attention heads specializing in *universal* commonsense reasoning. This reveals new linguistic properties of masked language models trained on multiple languages.

2 Related work

2.1 Winograd Schema challenges

The Winograd Schema Challenge (WSC) was proposed as a challenging yet practical benchmark for evaluation of machine commonsense reasoning (Levesque et al., 2012). Since its introduction, several English-language benchmarks of varying difficulty and size were also proposed: notable examples include Definite Pronoun Resolution (Rahman and Ng, 2012) and Pronoun Disambiguation Problem (Morgenstern et al., 2016) datasets, as well as WinoGrande, which consists of 44k crowdsourced examples (Sakaguchi et al., 2020). A version of WSC is also included in the popular SuperGLUE language understanding benchmark (Wang et al., 2019a), where it is reformulated as a natural language inference problem.

There also exist variations of WSC in other languages: French (Amsili and Seminck, 2017), Japanese (Shibata et al., 2015), Russian (Shavrina et al., 2020), Portuguese (Melo et al., 2019), and Chinese (Bernard and Han, 2020). We use these datasets in our study to create a multilingual dataset for commonsense reasoning.

Although in general the task definition of Winograd Schema Challenge was formalized to some degree, both succeeding datasets and methods proposed by users of these datasets have introduced various changes to the task specification and even the input format. In particular, a work by Liu et al. (2020) provides a thorough comparison of different ways to formalize the task for WSC and shows that the same model can give widely varying results depending on the evaluation framework. We describe our efforts to convert different datasets to a single format in Section 4.

2.2 Language models applied to commonsense reasoning

Several works attempt to solve Winograd Schema Challenge by utilizing pretrained language models. For example, Trinh and Le (2018) propose to rank possible answers with an ensemble of RNN language models by substituting the pronoun with each of the options. Recently, Klein and Nabi (2019) introduced Maximum Attention Score (MAS) for commonsense reasoning. This method uses the outputs of multi-head attention from each layer and scores each candidate answer based on the number of heads for which this answer has the highest attention value. We use the first (adapted to

[1] Our datasets and code are available at github.com/yandex-research/crosslingual_winograd

masked language models as proposed by Salazar et al., 2020) and the second approaches as baselines in the experiments. In essence, our method can be compared to MAS, but as we demonstrate in Section 5, several algorithm design differences along with task supervision allow us to significantly improve the commonsense reasoning performance.

Large pretrained Transformer models, such as BERT (Devlin et al., 2019), have also enabled rapid progress of supervised methods for WSC. One such method is given by Sakaguchi et al. (2020): the authors propose to concatenate the sentence and one of the options and to use the [CLS] token representation of the resulting sequence for binary classification. Also, Kocijan et al. (2019) propose a margin-based loss function which aims to increase the log-probability of the correct answer as a replacement for the masked pronoun. We evaluate these methods in our experiments without training on large in-domain datasets; as we show, both methods are prone to overfitting when applied to several hundreds of examples.

2.3 Cross-lingual encoder models

Multilingual representations have been a long-standing goal of the research community: they allow to serve fewer models for a wide range of languages and to improve the results on low-resource languages. Ruder et al. (2019) gives a detailed survey of different cross-lingual word embedding approaches, as well as the history of cross-lingual representations in general.

In this work, we are interested in the latest developments in multilingual Transformer masked language models (Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020; Siddhant et al., 2020) that were driven by the advances in transfer learning for NLP (Howard and Ruder, 2018; Devlin et al., 2019). In particular, we use pretrained multilingual BERT (mBERT, Devlin et al., 2019) and XLM-RoBERTa (XLM-R, Conneau et al., 2020) for all our experiments.

Recently, there has been increasing interest in the evaluation of multilingual models: as a result, several benchmarks, including XTREME (Hu et al., 2020), XNLI (Conneau et al., 2018) and XCOPA (Ponti et al., 2020) were introduced. Although XCOPA is a commonsense reasoning dataset, it is meant to serve as a multilingual version of the COPA dataset (Roemmele et al., 2011), which offers a problem different from pronoun resolution. In this work, we aimed to create a multilingual counterpart of more widely used Winograd Schema Challenge, so that any future methods for commonsense reasoning can be easily evaluated on languages other than English.

2.4 Functions of Transformer heads

Previous works have demonstrated that it is possible to perform unsupervised zero-shot consistency parsing with attention heads of pretrained cross-lingual models (Kim et al., 2020; Li et al., 2020). In our work, we extend these findings to a conceptually different task of commonsense reasoning. This task has significant overlap with coreference resolution, which was shown to be encoded in specific heads of monolingual BERT (Clark et al., 2019; Tenney et al., 2019).

Motivated by similar results for monolingual models, several works have previously demonstrated that models such as multilingual BERT encode grammatical relations (Chi et al., 2020) and can perform zero-shot entity recognition, as well as POS-tagging (Pires et al., 2019). Besides presenting evidence for universality in pronoun resolution, which was not studied before, our analysis relies on attention heads instead of extracting representations from intermediate layer outputs.

3 Common sense from attention

In this section, we first give a formal definition of the commonsense reasoning task, most commonly encountered in Winograd Schema Challenge and its successors. Then, we provide necessary background information about the Transformer architecture for transfer learning and describe our proposed solution for this task.

3.1 Exact task specification

It is known that commonsense reasoning performance can vary greatly due to changes in task formulation: for example, recent work by Liu et al. (2020) reports improvements of up to 6 points when posing the task as multiple choice instead of binary classification. Thus, as per recommendations from this work and in order to create a unified dataset, we choose the definition of the Winograd Schema problem which is as strict as possible.

The definition is as follows: the system receives a sentence with a pronoun and has to choose the noun (or noun phrase) that this pronoun refers to. For this choice, the system has two options; both of which, along with the pronoun, are always included as substrings of the initial sentence. We intentionally do not restrict the choice of sentence representation or the framing of the task in order to evaluate a diverse range of solutions.

Although the requirements listed above are quite general and intuitive when working with WSC, some of the datasets we employ have samples that do not conform to them. For example, it might be the case that the pronoun occurs at several positions in the sentence without explicit indication of the one to be resolved. For all such examples, we attempt to convert them to standardized instances by hand and drop them only if it is not possible via simple means: otherwise, the right answer to the problem is misspecified. We give a detailed description of our solution in Section 4.2.

3.2 Transformers for sentence representations

Our method heavily relies on the specifics of the Transformer architecture (Vaswani et al., 2017), which has attracted increased interest in NLP recently due to its generation (Raffel et al., 2020; Brown et al., 2020) and transfer learning (Devlin et al., 2019; Liu et al., 2019) capabilities.

This architecture consists of several sequential layers, where each layer contains a feed-forward block and a self-attention block. Inside the self-attention block, there are multiple *attention heads*: each head first linearly projects the input sequence $z = [z_1, \dots, z_i, \dots, z_n]$ into sequences of queries q_i, keys k_i and values v_i, then computes the attention weights as softmax-normalized values of pairwise dot products between all keys and all queries:

$$\alpha_{ij} = \frac{\exp(q_i^T k_j)}{\sum_{l=1}^{n} \exp(q_i^T k_l)} \quad (1)$$

These weights are then used to combine the values into a single vector for each input vector, and the layer output is a linear combination of all attention head outputs.

3.3 Our approach

The method proposed in this work uses intermediate outputs of a Transformer masked language model with L layers and H heads in each layer. Given an instance of the Winograd Schema problem, we take the input sentence and mask the pronoun that needs to be resolved. After that, we feed the resulting sentence to the language model and obtain the activations of each self-attention layer as a tensor $L \times H \times T$, where T is the number of tokens that constitute the candidate answer. Here, we can either take the attention from the pronoun to the candidate or vice versa.

After aggregating the attention outputs by computing the mean or the maximum over T, we have two matrices for each of two possible answers, which are then flattened into vectors. Combining these vectors, we obtain an input for the binary classification task with class 0 corresponding to the first answer being correct and class 1 corresponds to the second one. Given a dataset of such inputs, we can train a logistic regression to predict the class from the multi-head attention weights α.

There are several design choices which define the exact implementation of our method. We describe them below; for each design choice, we underline the best-performing option as found by the ablation study in Section 5.4.

Feature combination: With two feature vectors for candidate answers, we can either concatenate them or subtract the vector of the second candidate from the vector of the first one.

Pooling over tokens: As the candidates can have different length, we need to transform the attention outputs to feature vectors of the same size. This can be done by one of two simple forms of aggregation: mean- or max-pooling.

Attention direction: Observe from Equation 1 that in general, $\alpha_{ij} \neq \alpha_{ji}$. To find the optimal configuration, we evaluate both options of either attending to the candidate or the pronoun.

4 Dataset

In this section, we describe our procedure of building a multilingual commonsense reasoning benchmark using Winograd Schema Challenge problems. We create this benchmark by combining several monolingual collections for six languages, each described in previously published works.

We intentionally do not use XCOPA (Ponti et al., 2020) as it is aimed at a different problem: instead of operating at the word level, the task of this dataset is to connect the premise and one of two hypotheses, both of which are complete sentences. Because direct application of attention-based reasoning to sentence-level tasks is a non-trivial research question, we leave it to future work.

4.1 Languages

For the English language, we work with the data from the original WSC task[2] (Levesque et al., 2012), as well as the SuperGLUE benchmark (Wang et al., 2019a) and the Definite Pronoun Resolution dataset (Rahman and Ng, 2012). For French and Japanese, we use datasets published by Amsili and Seminck (2017) and Shibata et al. (2015) respectively. We also include the corresponding part from the Russian SuperGLUE benchmark (Shavrina et al., 2020), a collection of Winograd Schemas in Chinese from the WSC website[3], and the Portuguese version of WSC (Melo et al., 2019) into our multilingual benchmark.

In addition, we attempted to use Mandarinograd (Bernard and Han, 2020) — a Mandarin Chinese version of WSC. However, this dataset contains questions instead of pronouns that need to be resolved. As such, we were unable to incorporate its contents without significantly changing the task.

4.2 Preprocessing and filtering

As the datasets for different languages were released in several different formats, in order to have a unified evaluation framework, we needed to convert them all to the same schema. Unfortunately, due to the differences in task formalization we were unable to convert certain examples without completely changing them; as a result, these examples had to be removed from the dataset. Still, our main priority was to maintain the same task format while keeping as many examples as possible; to this end, we fixed minor annotation inconsistencies by hand wherever possible.

Below we describe the steps of our pipeline. First, several examples had more than two candidate choices, i.e. more than one incorrect option is given. We convert these examples into several binary choice problems and report the original dataset sizes after executing this step. Next, the main issue we faced was that the right answer is not included as a substring of the input sentence. Often this can be explained by missing articles, typos or differences in word capitalization. We attempt to fix all such errors in these cases.

The resulting dataset sizes are listed in Table 1; it can be seen that our conversion pipeline discards approximately 29% of data. In the future, more effort could be directed towards constructing a linguistically diverse, large-scale and balanced multilingual Winograd Schema dataset. Yet, as shown in Section 5, this collection of datasets already allows us to distinguish recent commonsense reasoning models by their performance.

Language	Before	After	Remaining, %
English	2605	2325	89.25
French	214	83	38.79
Japanese	1886	959	50.85
Russian	569	315	55.36
Chinese	18	16	92.28
Portuguese	285	263	88.89
Total	5577	3961	71.02

Table 1: Dataset sizes before and after filtering.

5 Experiments

Below we describe the experimental setup used to evaluate cross-lingual transfer capabilities of different approaches to commonsense reasoning and report the results. Note that we also aim to study the universal reasoning properties of attention heads, and thus we do not evaluate our method on common monolingual Winograd Schema datasets.

5.1 Setup

Models We use multilingual BERT (Devlin et al., 2019) and XLM-R-Large (Conneau et al., 2020), as these models are frequently used in other multilingual evaluation literature. The first model has 12 layers with 12 attention heads each, whereas the second model is a 24-layer Transformer with 16 attention heads on each layer. We do not evaluate XLM-R-Base or multilingual translation encoders (Siddhant et al., 2020) because we take two best-performing models according to the XTREME benchmark (Hu et al., 2020).

For our method, we use an implementation of logistic regression from scikit-learn (Pedregosa et al., 2011) with default hyperparameters as a linear classifier over attention weights.

Evaluation For unsupervised methods, we directly apply each method to each language subset and report the classification accuracy. For supervised methods, we first choose a single language for training and generate random train-validation-test splits, leaving 10% of data both for validation and testing subsets. For each language, we create 5

[2] Specifically, the WSC285 version.

[3] https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WSChinese.html

Model	Train lang	en	fr	ja	ru	zh	pt	Avg
Unsupervised								
MLM prob. ranking	-	53.6	53.0	52.5	51.8	31.3	50.2	**52.8**
Pseudo-perplexity	-	53.0	54.2	49.5	53.7	56.3	49.4	52.0
MAS	-	52.3	51.8	50.2	52.7	56.3	49.1	51.6
Supervised								
Kocijan et al. (2019)	en	-	52.5±4.1	51.4±0.8	51.2±1.1	48.8±9.3	51.2±1.5	51.0
	fr	50.9±0.4	-	51.5±0.7	51.3±0.9	56.2±9.9	49.2±1.2	51.8
	ja	51.0±0.7	50.8±2.2	-	50.1±1.0	55.0±6.8	49.9±1.1	51.4
	ru	50.7±0.5	51.3±2.0	51.7±1.0	-	51.2±10.3	51.0±0.3	51.2
	zh	50.9±0.2	50.1±1.6	50.8±0.4	50.5±0.0	-	53.0±1.4	**51.1**
	pt	51.1±0.5	54.0±3.9	51.3±0.6	49.3±0.5	53.8±7.1	-	51.9
Ours	en	-	53.7±1.6	52.2±0.4	60.1±0.4	51.2±4.7	53.9±0.4	**54.2**
	fr	51.7±1.1	-	51.1±1.1	52.2±2.6	53.8±3.1	50.9±1.1	**51.9**
	ja	52.7±0.3	55.4±0.8	-	58.0±1.2	50.0±4.0	51.3±1.3	**53.5**
	ru	55.5±0.4	52.3±2.7	52.3±0.3	-	52.5±5.0	52.0±0.7	**52.9**
	zh	49.8±2.0	48.2±4.5	50.5±1.4	50.3±4.9	-	49.0±1.4	49.6
	pt	54.7±0.5	52.8±3.2	51.7±0.6	57.7±1.2	50.0±4.0	-	**53.4**

Table 2: Results for multilingual BERT, best result is denoted by bold font.

random train-validation splits to estimate the standard deviation of metrics, while keeping the same test set to keep the results comparable. Additionally, we test each trained model for a language on all other languages in a zero-shot setting, reporting averaged performance as well.

5.2 Baselines

To compare our approach with currently popular methods, we also evaluate a wide set of well-performing approaches described in earlier works:

Unsupervised We use three entirely unsupervised baselines inspired by prior work. For the first approach, we replace the pronoun by the number of `[MASK]` tokens equal to the length of each candidate answer and compare the MLM probabilities. For the second approach, we replace the pronoun with each of the answers and rank the candidates by "pseudo-perplexity" (Salazar et al., 2020), inspired by the results of Trinh and Le (2018). Both baselines use normalized scores with respect to the candidate word length.

The third unsupervised baseline is Masked Attention Score (MAS), described in Klein and Nabi (2019). Similarly to our method, this approach relies on attention weights for prediction; however, they are utilized differently and the model is unable to discover an optimal subset of heads.

Supervised First, we evaluated the masked language model finetuning approach suggested by the authors of WinoGrande (Sakaguchi et al., 2020). However, in our experiments there are no additional large-scale datasets; we found that with reference hyperparameters, the authors' implementation quickly overfits the training data for all languages in our relatively small benchmark, achieving less than 50% zero-shot accuracy on average.

In addition, we used the margin-based classification approach described in (Kocijan et al., 2019). This method achieves competitive results and outperforms unsupervised baselines in most setups, so we include it in our comparison.

5.3 Results

The results of our experiments for multilinual BERT and XLM-R-Large are shown in Tables 2 and 3 respectively. It can be seen that despite using only the attention weights as features, our method can outperform unsupervised approaches and performs competitively with a state-of-the-art supervised approach in several setups. Notably, the quality improves significantly when going from BERT to XLM-R: this goes in line with previous work on evaluation of cross-lingual encoders (Hu et al., 2020). At the same time, the quality of our method improves more significantly than of that suggested by Kocijan et al. (2019): this may be explained by a greater parameter count and a higher number of attention heads with more distinct specializations.

5.4 Ablation study

Here we compare several algorithm versions listed in Section 3.3. We train all models on the English part of the dataset and evaluate on all other languages, using validation subset performance as our

Model	Train lang	en	fr	ja	ru	zh	pt	Avg
				Unsupervised				
MLM prob. ranking	-	58.8	56.6	61.7	57.5	56.3	56.7	**59.2**
Pseudo-perplexity	-	58.5	54.2	58.2	59.7	56.3	54.8	58.2
MAS	-	57.2	56.6	53.9	58.1	50.0	53.6	56.2
				Supervised				
Kocijan et al. (2019)	en	-	70.6±2.0	81.4±1.8	74.8±1.1	72.5±5.6	74.3±1.2	**74.7**
	fr	59.6±0.5	-	65.8±0.2	58.9±0.3	56.2±0.0	56.7±0.9	59.4
	ja	70.4±3.8	62.7±2.7	-	66.1±1.9	63.7±5.2	63.7±2.7	65.3
	ru	67.1±4.8	62.2±4.2	69.2±3.2	-	56.2±0.0	61.0±4.3	63.1
	zh	59.2±0.0	57.8±0.0	65.5±0.0	58.7±0.0	-	56.3±0.0	**59.5**
	pt	67.1±3.6	61.2±2.2	69.4±3.1	65.6±3.7	60.0±3.4	-	64.6
Ours	en	-	67.5±1.3	69.1±0.2	70.4±0.5	60.0±3.1	66.8±0.9	66.7
	fr	66.1±0.7	-	63.3±0.8	67.0±1.2	60.0±5.0	61.4±1.2	**63.6**
	ja	70.1±0.4	67.2±1.4	-	72.4±0.6	61.3±2.5	65.9±0.8	**67.4**
	ru	68.7±0.5	65.8±2.7	65.7±0.5	-	63.7±4.7	64.4±0.8	**65.7**
	zh	51.1±8.5	48.2±8.6	52.4±8.8	51.3±10.8	-	50.0±8.1	50.6
	pt	68.9±0.6	67.0±1.6	68.1±0.4	69.5±0.2	63.7±4.7	-	**67.4**

Table 3: Results for XLM-R-Large, best result is denoted by bold font.

Method	Valid	fr	ja	ru	zh	pt	Avg
Ours (Section 3.3)	**55.4**	53.7	52.2	**60.1**	51.2	**53.9**	**54.2**
Concat	53.0	**54.9**	**52.3**	56.3	**53.8**	**53.9**	**54.2**
Max pooling	53.6	**52.3**	52.3	59.9	50.0	51.6	53.2
Attn from pronoun	54.8	53.0	52.2	57.4	47.5	52.9	52.6

Table 4: Ablation study results for models trained on the English subset; the best result is in bold.

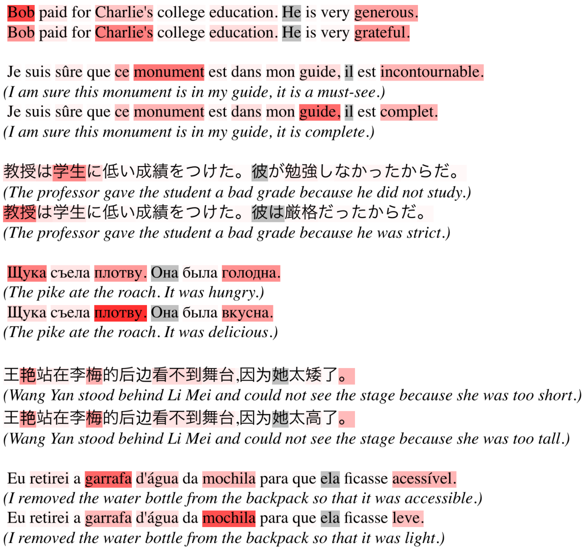
target metric. As the Table 4 demonstrates, each choice leads to drops in performance, with the most influential being the choice of feature concatenation instead of taking the difference and attention direction being the least important decision.

6 Analyzing the attention heads

In this section, we intend to analyze the reasons behind competitive generalization performance of our approach. Mainly we compare the subsets of heads learned on different languages and measure their impact on the prediction quality.

6.1 Universal commonsense reasoning

For the first experiment, we rank the heads for models trained on all languages with the XLM-R[4] representations by the absolute value of the weight. Then, we consider the top-5 heads which are ranked highest on average across all languages. These common heads are located in the higher layers of the model, which was shown previously to encode mainly semantic features (Raganato and Tiedemann, 2018; Jo and Myaeng, 2020), which intuitively corresponds to the tasks the model needs to solve for pronoun resolution. Figure 2 shows the average attention weights of these heads for each word in several example sentences.



Bob paid for Charlie's college education. He is very generous.
Bob paid for Charlie's college education. He is very grateful.

Je suis sûre que ce monument est dans mon guide, il est incontournable.
(I am sure this monument is in my guide, it is a must-see.)
Je suis sûre que ce monument est dans mon guide, il est complet.
(I am sure this monument is in my guide, it is complete.)

教授は学生に低い成績をつけた。彼が勉強しなかったからだ。
(The professor gave the student a bad grade because he did not study.)
教授は学生に低い成績をつけた。彼は厳格だったからだ。
(The professor gave the student a bad grade because he was strict.)

Щука съела плотву. Она была голодна.
(The pike ate the roach. It was hungry.)
Щука съела плотву. Она была вкусна.
(The pike ate the roach. It was delicious.)

王艳站在李梅的后边看不到舞台,因为她太矮了。
(Wang Yan stood behind Li Mei and could not see the stage because she was too short.)
王艳站在李梅的后边看不到舞台,因为她太高了。
(Wang Yan stood behind Li Mei and could not see the stage because she was too tall.)

Eu retirei a garrafa d'água da mochila para que ela ficasse acessível.
(I removed the water bottle from the backpack so that it was accessible.)
Eu retirei a garrafa d'água da mochila para que ela ficasse leve.
(I removed the water bottle from the backpack so that it was light.)

Figure 2: Averaged attention from the pronoun when using top-5 common heads.

After we locate the most important common heads, we train linear classifiers restricted to these heads as features only for every language. To evaluate the importance of head choice, we also provide the performance of linear classifiers trained on a fixed subset of 5 random heads. The results of this experiment can be seen in Table 5; we observe that using the same top-5 heads (only 1.3% of the total number) across all languages preserves or even improves the results. The only exception is Chinese, which might not have enough labeled data to extract a sufficient amount of task-specific in-

[4] The results for mBERT are available in Appendix B.

Train lang	Heads	en	fr	ja	ru	zh	pt	Avg
				MAS (unsupervised)				
-	All	57.2	56.6	53.9	58.1	50.0	53.6	56.2
	Random	57.8	56.6	56.9	61.6	50.0	56.7	56.6
	Common	**65.8**	**62.7**	**64.9**	**67.3**	**68.8**	**64.3**	**65.6**
				Ours (supervised)				
en	All	-	67.5	**69.1**	**70.4**	60.0	**66.8**	**66.7**
	Random		62.0	64.4	67.4	60.0	65.4	63.9
	Common		**68.4**	66.6	68.5	62.5	65.3	66.3
fr	All	66.1	-	63.3	**67.0**	60.0	61.4	63.6
	Random	59.9		58.3	60.7	58.8	57.2	59.0
	Common	**66.7**		**63.8**	66.7	**63.7**	**63.1**	**64.8**
ja	All	**70.1**	**67.2**	-	**72.4**	61.3	**65.9**	**67.4**
	Random	66.0	62.2		68.0	59.4	65.3	64.2
	Common	68.9	66.7		69.5	**62.5**	64.9	66.5
ru	All	**68.7**	**65.8**	65.7	-	**63.7**	64.4	**65.7**
	Random	66.0	62.3	64.3		59.4	**64.6**	63.3
	Common	68.0	64.6	**66.5**		**63.7**	**64.6**	65.5
zh	All	51.1	48.2	52.4	51.3	-	50.0	50.6
	Random	**59.4**	**54.7**	**58.6**	**61.0**		**58.0**	**58.3**
	Common	46.4	47.2	49.4	46.8		46.9	47.4
pt	All	**68.9**	**67.0**	**68.1**	**69.5**	**63.7**	-	**67.4**
	Random	66.2	62.3	64.6	67.1	60.0		64.0
	Common	67.9	65.5	66.0	68.2	**63.7**		66.3

Table 5: Performance of models trained with different subsets of XLM-R-Large attention heads.

formation. It means that a very small subset of attention weights is required to perform commonsense reasoning in all evaluated languages. This further supports the previous results on the analysis of linguistic universals in cross-lingual models (Chi et al., 2020; Wang et al., 2019b).

Moreover, restricting the subset of heads used in the MAS baseline to those selected by the classifiers significantly improves the quality of this unsupervised method as well, nearly closing the gap with the results obtained with supervision. This leads us to the conclusion that initially the poor performance of MAS might be caused by the suboptimal choice of attention heads; when the right heads are selected, their weights do not impact the predictions as significantly. Future unsupervised methods for commonsense reasoning can use that information to pay more attention to the choice of heads, which is currently a less explored subject.

6.2 The impact of number of heads

In this experiment, we directly study the connection between the number of heads and the quality of predictions. Specifically, after training a model with a full set of attention heads, we order them by the absolute value. Then, we retrain the model while keeping only the top-N important heads.

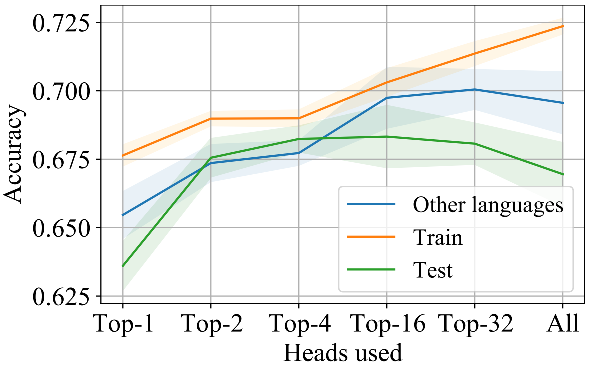


Figure 3: Effect of the number of XLM-R attention heads used when training on English data. Shaded areas show standard deviation across runs.

Figure 3 displays the results of our study for the English language; results for other languages can be seen in Appendix C. From these results, we find that although the training accuracy monotonically increases with the number of used attention weights, the optimal amount of heads for cross-lingual generalization is approximately equal to 16. This number is optimal or near-optimal for other languages as well, which might mean that as the number of features grows, the model either simply overfits the data or starts relying on features that are not universal for all languages.

7 Conclusion

In this work, we offer a simple supervised method to utilize pretrained language models for commonsense reasoning. It relies only on the outputs of self-attention and outperforms complete finetuning in a zero-shot scenario.

We also create a multilingual dataset of Winograd schemas that contains tasks from English, French, Japanese, Russian, Chinese, and Portuguese languages with the same specification. We want to encourage research on commonsense reasoning in languages other than English and release our benchmark to facilitate the development and analysis of new methods for this problem.

Lastly, we demonstrate that the reasoning capabilities of cross-lingual models are concentrated in a small subset of attention heads located in higher layers of the model. Furthermore, this subset of heads is language-agnostic, which sheds light at another facet of linguistic universals encoded in models such as multilingual BERT and XLM-R.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot.

Pascal Amsili and Olga Seminck. 2017. A Google-proof collection of French Winograd schemas. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 24–29, Valencia, Spain. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Timothée Bernard and Ting Han. 2020. Mandarinograd: A Chinese collection of Winograd schemas. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 21–26, Marseille, France. European Language Resources Association.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

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.

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.

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.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 328–339.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Jae-young Jo and Sung-Hyon Myaeng. 2020. Roles and utilization of attention heads in transformer-based neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3404–3417, Online. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

Taeuk Kim, Bowen Li, and Sang goo Lee. 2020. Chart-based zero-shot constituency parsing on multiple languages.

Tassilo Klein and Moin Nabi. 2019. Attention is (not) all you need for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4831–4836, Florence, Italy. Association for Computational Linguistics.

Tassilo Klein and Moin Nabi. 2020. Contrastive self-supervised learning for commonsense reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7517–7523, Online. Association for Computational Linguistics.

Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz. 2019. A surprisingly robust trick for the Winograd schema challenge. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4837–4842, Florence, Italy. Association for Computational Linguistics.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Citeseer.

Bowen Li, Taeuk Kim, Reinald Kim Amplayo, and Frank Keller. 2020. Heads-up! unsupervised constituency parsing via self-attention heads. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 409–424, Suzhou, China. Association for Computational Linguistics.

Haokun Liu, William Huang, Dhara Mungra, and Samuel R. Bowman. 2020. Precise task formalization matters in Winograd schema evaluations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8275–8280, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Gabriela Melo, Vinicius Imaizumi, and Fábio Cozman. 2019. Winograd schemas in portuguese. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*, pages 787–798. SBC.

Leora Morgenstern, Ernest Davis, and Charles Ortiz. 2016. Planning, executing, and evaluating the winograd schema challenge. *AI Magazine*, 37:50–54.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Alessandro Raganato and Jörg Tiedemann. 2018. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297, Brussels, Belgium. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The Winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, Jeju Island, Korea. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8732–8740.

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

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.

Tomohide Shibata, Shotaro Obama, and Sadao Kurohashi. 2015. Building and analyzing the winograd schema challenge. *Proceedings of the Twenty-first Annual Meeting of the Association for Natural Language Processing*, 57(1):22–23.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2827–2835, Online. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Trieu H. Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019b. Cross-lingual BERT transformation for zero-shot dependency parsing. 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 5721–5727, Hong Kong, China. Association for Computational Linguistics.

A In-language metrics for supervised methods

Here, we provide the metrics of our method and the finetuning baseline described by Kocijan et al. (2019) that were obtained on the training, validation and test data for the same language that the models were trained. Tables 6 and 7 demonstrate the results: it can be seen that although our approach performs less well on the same language that was used for training, the issue of overfitting on train data is less noticeable, which might be the reason for better zero-shot metrics.

Model	Train lang	Train	Test
Kocijan et al. (2019)	en	100.0±0.0	47.8±2.3
	fr	100.0±0.0	44.4±7.9
	ja	100.0±0.0	46.9±1.0
	ru	100.0±0.0	52.5±1.4
	zh	100.0±0.0	20.0±44.7
	pt	100.0±0.0	49.6±6.2
Ours	en	57.5±0.3	52.7±1.1
	fr	54.5±3.2	33.3±17.2
	ja	54.1±0.6	49.6±3.1
	ru	60.8±0.9	46.2±2.3
	zh	61.7±8.5	20.0±24.5
	pt	57.1±0.4	43.0±6.9

Table 6: Train and test set metrics for supervised methods, multilingual BERT.

Model	Train lang	Train	Test
Kocijan et al. (2019)	en	100.0±0.0	83.3±1.0
	fr	100.0±0.0	44.4±11.1
	ja	100.0±0.0	79.6±2.2
	ru	100.0±0.0	60.0±3.4
	zh	100.0±0.0	50.0±0.0
	pt	100.0±0.0	55.6±5.2
Ours	en	71.6±0.4	67.2±1.2
	fr	67.7±1.4	37.8±5.4
	ja	71.2±0.2	65.6±1.7
	ru	71.4±0.9	58.1±1.5
	zh	73.3±3.3	20.0±24.5
	pt	67.1±0.8	68.1±5.0

Table 7: Train and test set metrics for supervised methods, XLM-R Large.

B Analysis of common heads for multilingual BERT

Table 8 shows the evaluation results of models using top-5 attention heads of multilingual BERT. It can be seen that leaving only 5 heads out of 144 improves average accuracy in all cases and per-language accuracy in 18/30 cases without any significant decreases in quality.

C Impact of number of heads for other languages

In this section, we analyze the changes in both supervised and zero-shot performance for our method that follow from changes in the number of used attention heads. Figure 4 displays the results for French, Japanese, Russian, and Portuguese language; we omit the results for the Chinese language due to high variance from the small training dataset size. From this figure, we observe the same trend: increasing the number of used heads past 16 can favorably affect the accuracy on the training set, but negatively impacts the resulting quality both for the test set and for other languages.

Train lang	Heads	en	fr	ja	ru	zh	Avg
				MAS (unsupervised)			
-	All	52.21	51.81	50.16	52.70	**56.25**	52.63
	Common	**56.60**	**53.01**	**51.82**	**60.00**	50.00	**54.29**
				Ours (supervised)			
en	All	-	53.33	52.05	58.92	52.88	54.29
	Common	-	**54.53**	**52.52**	**59.78**	52.25	**54.77**
fr	All	50.76	-	**50.41**	**51.80**	50.06	50.76
	Common	**51.01**	-	50.38	51.73	**50.62**	**50.94**
ja	All	53.25	**52.64**	-	57.48	**50.69**	53.51
	Common	**55.54**	51.84	-	**58.51**	50.56	**54.12**
ru	All	55.43	52.65	**52.00**	-	49.62	52.43
	Common	**56.20**	**52.92**	51.66	-	**49.75**	**52.63**
zh	All	50.28	50.12	50.09	50.53	-	50.25
	Common	**50.82**	**50.14**	**50.24**	**51.30**	-	**50.62**

Table 8: Performance of models trained with different sets of multilingual BERT attention heads.

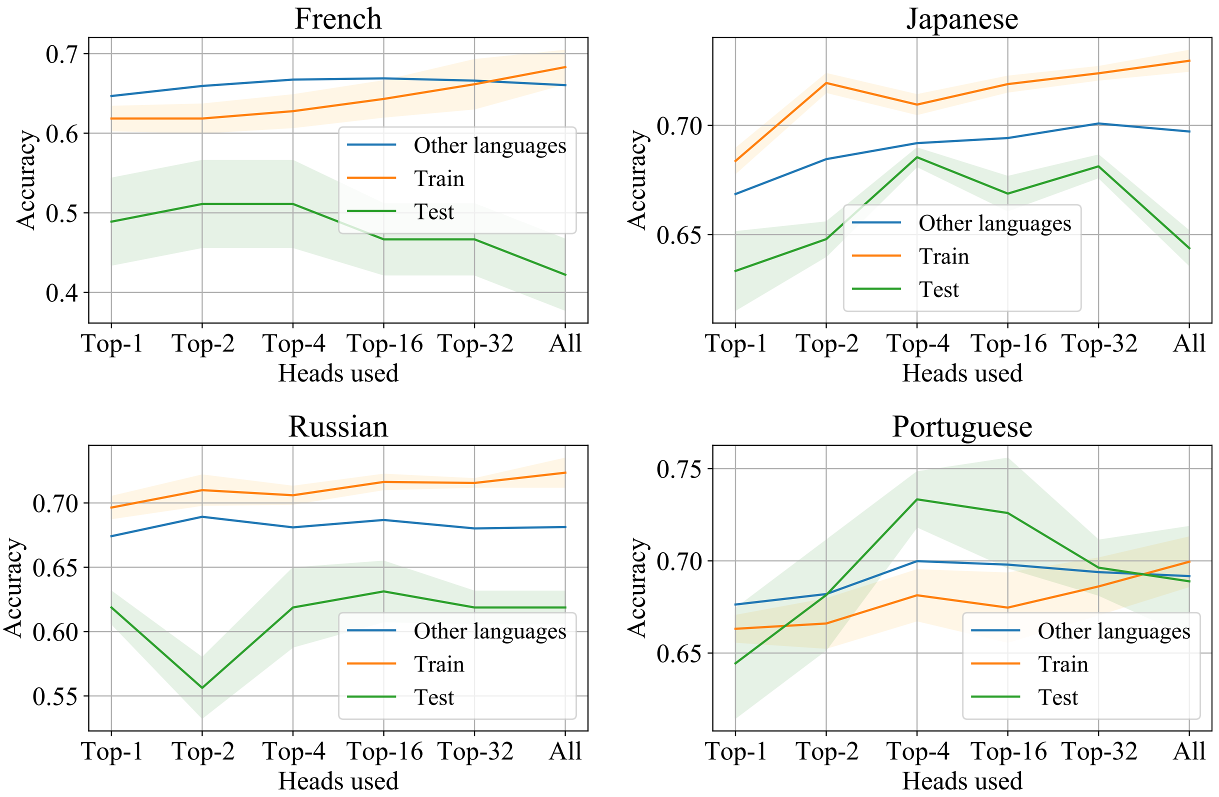


Figure 4: Effect of the number of used XLM-R attention heads on commonsense reasoning performance.