EXAMS: A Multi-Subject High School Examinations Dataset for Cross-Lingual and Multilingual Question Answering

Momchil Hardalov[1] **Todor Mihaylov**[2] **Dimitrina Zlatkova**[1] **Yoan Dinkov**[1]
Ivan Koychev[1] **Preslav Nakov**[3]
[1]Sofia University "St. Kliment Ohridski", Bulgaria,
[2]Heidelberg University, Germany
[3]Qatar Computing Research Institute, HBKU, Qatar,
{hardalov, koychev}@fmi.uni-sofia.bg
{dvzlatkova, jdinkov}@uni-sofia.bg
tbmihaylov@gmail.com, pnakov@hbku.edu.qa

Abstract

We propose $E\chi\alpha\mu s$ – a new benchmark dataset for cross-lingual and multilingual question answering for high school examinations. We collected more than 24,000 high-quality high school exam questions in 16 languages, covering 8 language families and 24 school subjects from Natural Sciences and Social Sciences, among others.

$E\chi\alpha\mu s$ offers a fine-grained evaluation framework across multiple languages and subjects, which allows precise analysis and comparison of various models. We perform various experiments with existing top-performing multilingual pre-trained models and we show that $E\chi\alpha\mu s$ offers multiple challenges that require multilingual knowledge and reasoning in multiple domains. We hope that $E\chi\alpha\mu s$ will enable researchers to explore challenging reasoning and knowledge transfer methods and pre-trained models for school question answering in various languages which was not possible before. The data, code, pre-trained models, and evaluation are available at http://github.com/mhardalov/exams-qa.

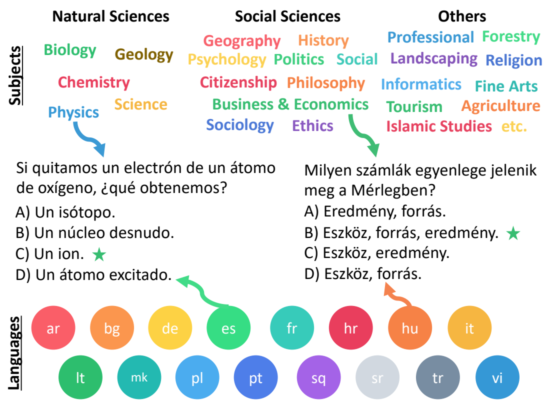

Figure 1: Properties and examples from $E\chi\alpha\mu s$.

1 Introduction

Research on science question answering has attracted a lot of attention in recent years (Clark, 2015; Schoenick et al., 2017; Clark et al., 2019). Such questions are challenging as they require domain and common sense knowledge (Clark et al., 2018), as well as complex reasoning and different forms of inference over a variety of knowledge sources (Khashabi et al., 2016, 2018). Indeed, a combination of these was required to achieve noticeable performance gains (Clark et al., 2016). This inevitably made research in school-level science Question Answering (QA) hard for languages other than English due to the scarceness of resources (Clark et al., 2014; Khot et al., 2017, 2018; Bhakthavatsalam et al., 2020).

There has been a recent mini-revolution in QA, as well as in the field of Natural Language Processing (NLP) in general, due to the invention of the Transformer (Vaswani et al., 2017), and the subsequent rise of large-scale pre-trained models (Peters et al., 2018; Radford et al., 2018, 2019; Devlin et al., 2019; Lan et al., 2020; Yang et al., 2019; Liu et al., 2019c; Raffel et al., 2020). Nowadays, fine-tuning such models on task-specific data has become an essential element of any top-scoring QA system. Yet, for science QA, training on datasets from a different domain (Sun et al., 2019; Khashabi et al., 2020) and carefully selected background knowledge (Banerjee et al., 2019; Ni et al., 2019) could improve such models further.

The success of large-scale pre-trained models and the development of their multilingual versions (Devlin et al., 2019; Conneau et al., 2020) gives hopes for supposedly better performance in multilingual question answering. Therefore, several new datasets have been released for multilingual reading comprehension and open-domain question answering in the Wikipedia domain (Liu et al., 2019a; Lewis et al., 2020; Artetxe et al., 2020; Clark et al., 2020).

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 5427–5444,
November 16–20, 2020.

Here, we present $E\chi\alpha\mu s$, a new dataset and benchmark for multilingual and cross-lingual evaluation of models and methods for answering diverse school science questions (see Figure 1).

Our contributions are as follows:

- We advance the task of science Question Answering (QA) with multilingual and cross-lingual evaluations.
- We collect a new challenging dataset $E\chi\alpha\mu s$ from multilingual high school examinations, which offers several advantages over existing datasets: (*i*) it covers various domains, (*ii*) it is nearly three times larger than pre-existing Science QA datasets, (*iii*) it extends multi-lingual QA tasks to more languages, (*iv*) the questions are written by experts, rather than translated or crowdsourced, (*v*) the questions are harder since they are from matriculation exams rather than 4-8th grade.
- We use fine-grained evaluation – per subject and per language – which yields more precise comparison between models.
- We perform extensive experiments and analysis using top-performing multilingual models (mBERT, XLM-R), and we show that $E\chi\alpha\mu s$ offers several challenges that such models would need to overcome in the future, including multi-lingual and cross-lingual knowledge retrieval, aggregation, and reasoning, among others.

We release our code, pre-trained models and data for research purposes.[1]

2 Related Work

Science QA The work in Science Question Answering emerged in recent years with the development of several challenging datasets. The most notable is ARC (Clark et al., 2018), which is a QA reasoning challenge that contains both *Easy* and *Challenge* questions from 4th to 8th grade examinations in the *Natural Science* domain. As in $E\chi\alpha\mu s$, the questions in ARC are created by experts, albeit our dataset covers a wide variety of high school (8th-12th grade) subjects including but not limited to, Natural Sciences, Social Sciences, Applied Studies, Arts, Religion, etc. (see Section 3.2 for details). We provide definitions of the less known subjects in $E\chi\alpha\mu s$ in Appendix B.1.

[1]The $E\chi\alpha\mu s$ dataset and code are publicly available at `http://github.com/mhardalov/exams-qa`

The early versions of ARC (Clark, 2015; Schoenick et al., 2017) inspired several crowdsourced datasets: Welbl et al. (2017) proposed a scalable approach for crowdsourcing science questions given a set of basic supporting science facts. Dalvi et al. (2019) focused on specific phenomena including understanding science procedural texts, Mihaylov et al. (2018) and Khot et al. (2020) studied multi-step reasoning, given a set of science facts and commonsense knowledge, Tafjord et al. (2019), and Mitra et al. (2019) worked on reasoning about qualitative relationships, and declarative texts, among others. Unlike these English-only datasets, $E\chi\alpha\mu s$ offers questions in 16 languages. Moreover, it contains questions about multiple subjects, which are presumably harder as they were extracted mostly from matriculation examinations (8-12th grade). Finally, $E\chi\alpha\mu s$ contains over 24,000 questions, which is more than three times as many as in ARC.

Multilingual and Cross-lingual QA Recently, several QA datasets have been created that cover languages other than English, but still focusing on one such language. Gupta et al. (2018) proposed a parallel QA task for English and Hindi, Liu et al. (2019b) collected a bilingual cloze-style dataset in Chinese and English. Jing et al. (2019) crowdsourced parallel paragraphs from novels in Chinese and English. A few datasets investigated multiple-choice school QA (Hardalov et al., 2019; Van Nguyena et al., 2020), albeit in a limited domain, and for lower school grades (1st-5th). Other efforts focused on building bi-lingual datasets that are similar in spirit to SQuAD (Rajpurkar et al., 2016) – extractive reading comprehension over open-domain articles. Such datasets are collected by crowdsourcing questions, following a procedure similar to (Rajpurkar et al., 2016), in Russian (Efimov et al., 2020), Korean (Lim et al., 2019), French (d'Hoffschmidt et al., 2020), or by translating existing English QA pairs to Spanish (Carrino et al., 2020).

Recently, some multilingual datasets, were released to the public. MLQA (Lewis et al., 2020), and XQuAD (Artetxe et al., 2020) use translations by professionals and extend the monolingual SQuAD (Rajpurkar et al., 2016) to 7 and 11 languages, respectively, thus forming cross-lingual evaluation benchmarks. Clark et al. (2020) collected an entirely new dataset (TyDi QA) of questions in 11 typologically diverse languages.

Lang	Family	#Subjects	Question Len	Choice Len	#Choices	#Questions	Vocab
Albanian	Albanian	8	15.0	5.0	4.0	1,505	11,572
Arabic	Semitic	5	10.3	3.4	4.0	562	5,189
Bulgarian	Balto-Slavic	6	13.0	3.3	4.0	2,937	15,127
Croatian	Balto-Slavic	14	14.7	4.1	3.9	2,879	20,689
French	Romance	3	18.4	10.5	3.5	318	2,576
German	Germanic	5	18.3	9.1	3.5	577	4,664
Hungarian	Finno-Ugric	10	11.6	5.9	3.9	2,267	15,045
Italian	Romance	12	20.0	5.6	3.9	1,256	9,050
Lithuanian	Balto-Slavic	2	9.7	4.7	4.0	593	5,394
Macedonian	Balto-Slavic	8	13.4	4.5	4.0	2,075	13,114
Polish	Balto-Slavic	1	13.7	4.3	4.0	1,971	18,990
Portuguese	Romance	4	19.9	8.6	4.0	924	6,811
Serbian	Balto-Slavic	14	15.4	4.3	3.9	1,637	15,509
Spanish	Romance	2	23.0	10.2	3.2	235	2,130
Turkish	Turkic	8	19.5	4.6	4.4	1,964	22,069
Vietnamese	Austroasian	6	37.0	6.4	4.0	2,443	6,076
#Langs 16	#Families 8	24	17.19	5.08	3.96	24,143	158,942

Table 1: Statistics about $E\chi\alpha\mu s$. The average length of the question (*Question Len*) and the choices (*sChoice Len*) are measured in number of tokens, and the vocabulary size (*Vocab*) is measured in number of words.

The task was to ask a question, and then the shortest span answering it from a list of paragraphs was selected. As these datasets are complementary, rather than making each other obsolete, hereby the recently released XTREME (Hu et al., 2020) benchmark combined them in a joint task. $E\chi\alpha\mu s$ differs from the aforementioned multilingual benchmarks in several aspects. First, we extend the multilingual QA efforts to a different, more challenging domain (Clark et al., 2018). Second, our datasets support more languages. Next, the questions in $E\chi\alpha\mu s$ are written by educational experts rather than non-expert annotators, making the evaluation results comparable to a top-performing student. Finally, our fine-grained evaluation for different subjects, languages, and combinations thereof allows for in-depth analysis and comparison.

3 $E\chi\alpha\mu s$ Dataset

We introduce $E\chi\alpha\mu s$, a new benchmark dataset for multilingual and cross-lingual question answering from high school examinations. In this section, we present the properties of the dataset, and we give details about the process of data collection, preparation and normalization, as well as information about the data splits, and the parallel questions.

3.1 Dataset Statistics

We collected $E\chi\alpha\mu s$ from official state exams prepared by the ministries of education of various countries. These exams are taken by students graduating from high school, and often require knowledge learned through the entire course.

	de	es	fr	hr	hu	it	mk	sq	sr
de	-								
es	199	-							
fr	253	120	-						
hr	189	134	109	-					
hu	456	159	274	236	-				
it	30	9	15	1,214	99	-			
mk	0	0	0	0	0	0	-		
sq	0	0	0	0	0	0	1,403	-	
sr	40	25	20	1,564	104	1,002	0	0	-
tr	0	0	0	0	0	0	1,222	981	0

Table 2: Parallel questions for different language pairs.

The questions cover a large variety of subjects and material based on the country's education system. Moreover, we do not focus only on major school subjects such as Biology, Chemistry, Geography, History, and Physics, but we also cover highly-specialized ones such as Agriculture, Geology, Informatics, as well as some applied and profiled studies. These characteristics make the questions in the dataset of very high variety, and not easily solvable, due to the need for highly specialized knowledge. Next, we discuss the cross-lingual and the multilingual properties of our dataset.

Parallel Questions Some countries allow students to take official examinations in several languages. Such parallel examinations also exist in our dataset. In particular, there are 9,857 parallel question pairs spread across seven languages as shown in Table 2. The parallel pairs are coming from Croatia (Croatian, Serbian, Italian, Hungarian), Hungary (Hungarian, German, French, Spanish, Croatian, Serbian, Italian), and North Macedonia (Macedonian, Albanian, Turkish).

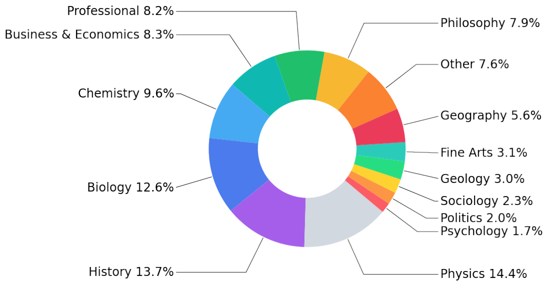

Figure 2: Relative sizes of the subjects. Those that cover less than 1.5% of the examples are in *Other*.

Multilinguality Our dataset includes a total of 24,143 questions in 16 languages from eight language families. Each question is a 3-way to 5-way (3.96 on average) multiple-choice question with a single correct answer. Table 1 shows a breakdown for each language, where the number of subjects, questions, and the vocabulary size are shown as absolute numbers, while the question length, the choice length, and the number of choices are averaged. All statistics about the questions and the answer options are measured in terms of words. We see that we have a rich vocabulary with almost 160,000 unique words. Interestingly, there are ∼9,500 shared words between at least one pair of languages in our dataset, excluding numbers and punctuation. As expected, the overlapping words are mostly between closely related languages (bg-mk, bg-sr, es-it, es-pt, hr-sr, mk-sr). Other common shared words are subject-specific words such as person names (e.g., *Abraham*, *Karl*, *Ivan*), chemical compounds (e.g., *NaOH*, *HCl*), units (e.g., *m/s*, *g/mol*), etc. Then, there are cognates with the exact same spelling (homographs) even between unrelated languages, mostly words of Latin or Greek origin, e.g., *temperatura* (temperature) and *forma* (form). Finally, there are also *false friends*, whose meaning differs across languages, e.g., *para* can mean *for* (es/pt) vs. *money* (mk/tr/sq) vs. *couple* (pl); similarly, *ser* can mean *be* (es/pt) vs. *cheese* (pl) vs. *after* (vi).

3.2 Subjects and Categories

Each education system has its own specifics, resulting in some differences in curricula, topics, and even naming of the subjects. That being said, the original, non-normalized categories in our dataset are more than 40 for exams from just a few countries. Given the sparse nature of the subjects, we use a two-level taxonomy in order to categorize them into logically connected groups. The lower-level is a subject, and the higher level is a major group. We normalized the subject using a two-step algorithm: first, we put each subject (with its original naming) in a separate category, then, if the subject was general enough, e.g., Biology, History, etc., or there were no similar ones, we retained the category; otherwise, we merged all similar subjects together in a unifying category, e.g., Economics Basics, and Economics & Marketing. We repeated the aforementioned steps until there were no suitable merge candidates. As a result, we ended up with a total of 24 subjects (see Appendix B for more details), which we further grouped into three major categories, based on the main branches of science: **Natural Science** – *"the study of natural phenomena"*, **Social Sciences** – *"the study of human behavior and societies"*, **Other** – *Applied Studies, Arts, Religion, etc.* (see Figure 1). [2]

The distribution of the major categories is *Natural Sciences* (40.0%) and *Social Sciences* (44.0%) and 16.0% for *Others* (these are the actual numbers, not approximate). The remaining questions are labeled as *Other* as they are not suitable for the two main categories. Figure 2 presents the relative sizes of the subjects in the dataset.

3.3 Collection and Preparation

Here, we describe the process of collecting and preparing the data, as it is not trivial and it could be applied to other languages and examinations. First, we identified potential online sources of publicly available school exams starting from the *Matriculation Examination* page in Wikipedia.[3]

For all languages in our dataset, the first step in the process of data collection was to download the PDF files per year, per subject, and per language (when parallel languages were available in the same source). We converted the PDF files to text and we used only those that were well-formatted and followed the document structure.

Then, we used Regular Expressions (RegEx) to parse the questions, their corresponding choices and the correct answer choice. In order to ensure that all our questions are answerable using textual input only, we removed questions that contained visual information. We did that using a manually curated list of words such as *map*, *table*, *picture*, *graph*, etc., in the corresponding language.

[2] https://en.wikipedia.org/wiki/Branches_of_science
[3] https://en.wikipedia.org/wiki/Matriculation_examination

Language	Multilingual Train	Dev	Test	Cross-lingual Train	Dev
Albanian	565	185	755	1,194	311
Arabic	-	-	562	-	-
Bulgarian	1,100	365	1,472	2,344	593
Croatian	1,003	335	1,541	2,341	538
French	-	-	318	-	-
German	-	-	577	-	-
Hungarian	707	263	1,297	1,731	536
Italian	464	156	636	1,010	246
Lithuanian	-	-	593	-	-
Macedonian	778	265	1,032	1,665	410
Polish	739	246	986	1,577	394
Portuguese	346	115	463	740	184
Serbian	596	197	844	1,323	314
Spanish	-	-	235	-	-
Turkish	747	240	977	1,571	393
Vietnamese	916	305	1,222	1,955	488
Combined	7,961	2,672	13,510	-	-

Table 3: Number of examples in the data splits based on the experimental setup.

Next, we performed data cleaning to ensure the quality of the generated dataset, by manually reviewing each question and its choices and ensuring that all options, text, and symbols (e.g., $\mu, \rightarrow, \alpha, \leftarrow$) were displayed correctly. As a result, we filtered out about 17% of the questions (the percentage varies based on the source, the language, and the subject). Finally, in order to remove frequency bias such as "most answers are B)", we shuffled each question's choices.

3.4 Data Splits

In our experiments, we aim at evaluating the multilingual and the cross-lingual question answering capabilities of different models. Therefore, we split the data in order to support both evaluation strategies: *Multilingual* and *Cross-lingual*.

Multilingual In this setup, we want to train and to evaluate a given model with multiple languages, and thus we need multilingual *training*, *validation* and *test* sets. In order to ensure that we include as many of the languages as possible, we first split the questions independently for each language L into Train_L, Dev_L, Test_L with 37.5%, 12.5%, 50% of the examples, respectively.[4] We then unite all language-specific subsets into the multilingual sets Train_{Mul}, Dev_{Mul}, Test_{Mul}, and we used them for training, development, and testing.

[4]For languages with fewer than 900 examples, we only have Test_L.

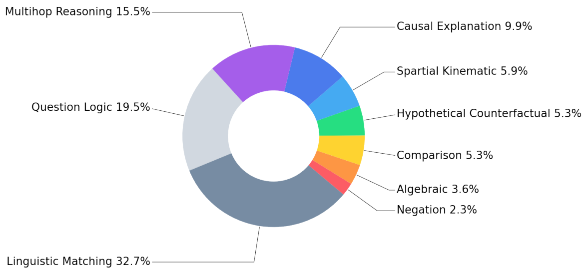

Figure 3: Relative sizes of reasoning types in $E\chi\alpha\mu s$.

Since we have parallel data for several languages (discussed in Section 3.1), in this setup, we ensure that the same parallel questions are only found in either training, development or testing, so that we do not leak the answer from training via some other language. In order to do that, we sample the questions with the assumptions and the ratios mentioned above, stratified per subject in the given language. The number of examples per language and the total number of multilingual sets are shown in the first three columns of Table 3.[5]

Cross-Lingual In this setting, we want to explore the capability of a model to transfer its knowledge from a single source language L_{src} to a new unseen target language L_{tgt}. In order to ensure that we have a larger training set, we train the model on 80% of L_{src}, we validate on 20% of the same language, and we test on a subset of L_{tgt}.[6] The last three columns of Table 3 show the number of examples used for training and validation with the corresponding language.

3.5 Reasoning and Knowledge Types

In order to give a better understanding of the reasoning, and the knowledge types in $E\chi\alpha\mu s$, we sampled and annotated 250 questions, all of which are from the multilingual Dev. For each question, we provided English translations as not all annotators were native speakers of the questions' language. We followed the procedure and re-used the annotation types presented in earlier work (Clark et al., 2018; Boratko et al., 2018). However, as they were designed mainly for Nature Science questions, we extended them with two new annotation types: *"Domain Facts and Knowledge"* and *"Negation"* (see Appendix C for examples).

[5]Sometimes, grouping parallel questions in the same split slightly violates the splitting ratios.

[6]To ensure that the cross-lingual evaluation is comparable to the multilingual one, we use the same subset of questions from language L_{tgt} that are used in Test_{Mul}

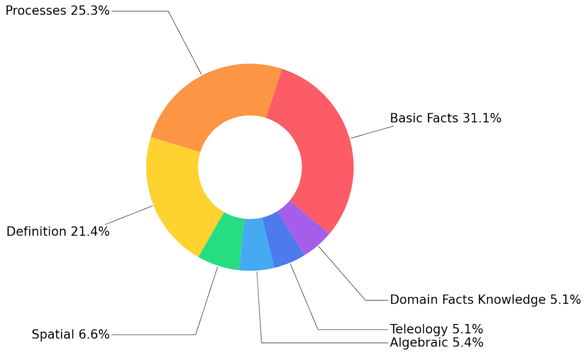

Figure 4: Relative size of the $E\chi\alpha\mu s$ knowledge types.

The relative sizes of the knowledge and the reasoning types are shown in Figures 3 and 4. Here, we must note that the sizes are approximate rather than exact, since the annotations are subjective and the distribution may vary.

4 Baseline Models

We divide our baselines into the following two categories: (*i*) models without additional training, and (*ii*) fine-tuned models. The first group contains common baselines, i.e., random guessing and information retrieval solver (Clark et al., 2016). In addition, we evaluate the knowledge contained in the pre-trained language model, i.e., mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020), and we use it as an answering mechanism. The second group of baselines compare the learning ability of state-of-the-art multilingual models on the task of multiple-choice question answering. Since we have multi-choice questions, we adopt accuracy as an evaluation measure, as this is standard for this setup.

4.1 No Additional Training

Information Retrieval (IR) This IR baseline is from Clark et al. (2016), and it ranks the possible options o for each question q based on the relevance score returned by a search engine[7]. In particular, for each option o_i, we form a query by appending the option's text to the question's $(q + o_i)$, and we send this concatenation to the search engine. We then sum the returned scores for the top-10 hits, and we predict the choice with the highest score to be the correct answer. More detailed discussion can be found in Appendix D.

[7]We build and use a separate index for each language using ElasticSearch.

Pre-trained Model as a Knowledge Base (KB) As we start to understand pre-trained BERT-like models better(Petroni et al., 2019; Rogers et al., 2020), we observe some interesting phenomena. Here, we evaluate the knowledge contained in the model by leveraging the standard masking mechanism used in pre-training. We tokenize each question-option pair into subwords, and then we replace all the pieces from the option with the special [MASK] token. Following the notation from Devlin et al. (2019), the input sequence can be written as follows:

[CLS] $[\mathrm{Q}_1] \ldots [\mathrm{Q}_\mathrm{N}]$ $[\mathrm{M\_O}_1] \ldots [\mathrm{M\_O}_\mathrm{M}]$ [SEP],

where Q is the question, and M_O is the masked option. Following the notation above, we obtain a score for each option in the question based on the normalized log-probability for the entire masked sequence. (see Eq. 1).

$$score(O_i) = \frac{1}{|O_i|} \sum_{t \in O_i} \log P_{MLM}(t|Q) \qquad (1)$$

We could probably obtain better results for that evaluation if we form the question-option pairs as a single statement, e.g., "What is the purpose of *something*? [SEP] [M_O] $\rightarrow$ The purpose of *something* is [M_O]."

4.2 Fine-Tuned Models

We are interested in evaluating the ability of pre-trained models to transfer science-based knowledge across languages when fine-tuned.

In order to evaluate the QA capability of these models, we follow the established approach in this setting (Devlin et al., 2019; Liu et al., 2019c; Sun et al., 2019), and we fine-tune them to predict the correct answer in a multi-choice setting, given a selected context. This setup feeds the pre-trained model with a text, tokenized using the corresponding tokenizer for the model in the format:

[CLS] C [SEP] Q + O [SEP],

where C, Q and O are the tokenized *knowledge context* (see Appendix D), the *question*, and the *option*, respectively. Each question-option pair (Q+O) is evaluated, and the one with the highest confidence of being an answer is selected.

In our experiments, we used the Transformers library (Wolf et al., 2019). We experimented with the best-performing multilingual models: the Multilingual version of BERT, or mBERT Devlin et al. (2019), and the recently proposed XLM-RoBERTa, or XLM-R (Conneau et al., 2020).

	ARC		R12	$E\chi\alpha\mu s$																
Lang/Set	**E**	**C**	**en**	**ar**	**bg**	**de**	**es**	**fr**	**hr**	**hu**	**it**	**lt**	**mk**	**pl**	**pt**	**sq**	**sr**	**tr**	**vi**	**All**
Random Guess	25.0	25.0	25.0	25.0	25.0	29.4	32.0	29.4	26.7	27.7	26.0	25.0	25.0	25.0	25.0	25.0	26.2	23.1	25.0	25.9
IR (Wikipedia)	-	-	-	31.0	29.6	29.3	27.2	32.1	31.9	29.7	27.6	29.8	32.2	29.2	27.5	25.3	31.8	28.5	27.5	29.5
XLM-R on RACE	61.6	45.9	57.4	39.1	43.9	37.2	40.0	37.4	38.8	39.9	36.9	40.5	45.9	33.9	37.4	42.3	35.6	37.1	35.9	39.1
w/ SciENs	**73.6**	51.2	68.4	39.1	44.2	35.5	37.9	37.1	38.5	37.9	39.5	**41.3**	49.8	36.1	**39.3**	42.5	37.4	37.4	35.9	39.6
then on $E\chi\alpha\mu s$ (Full)	72.8	**52.6**	**68.8**	**40.7**	**47.2**	**39.7**	**42.1**	**39.6**	**41.6**	**40.2**	**40.6**	40.6	**53.1**	**38.3**	38.9	**44.6**	**39.6**	**40.3**	**37.5**	**42.0**
XLM-R_{Base} (Full)	54.2	36.4	54.6	34.5	35.7	*36.7*	38.3	*36.5*	*35.6*	*33.3*	33.3	33.2	41.4	*30.8*	29.8	*33.5*	32.3	30.4	*32.1*	34.1
mBERT (Full)	*63.8*	*38.9*	*57.0*	34.5	*39.5*	35.3	*40.9*	34.9	35.3	32.7	*36.0*	*34.4*	*42.1*	30.0	29.8	30.9	*34.3*	*31.8*	31.7	*34.6*
mBERT ($E\chi\alpha\mu s$ only)	39.6	28.5	35.1	31.9	34.1	30.4	37.9	33.3	32.6	29.3	31.1	31.9	42.4	29.0	28.3	29.9	30.8	25.4	30.0	31.7
XLM-R as KB	30.8	26.2	27.2	31.0	27.2	31.7	37.9	29.9	27.6	29.3	28.0	28.3	23.5	24.6	27.0	25.6	25.4	24.4	24.9	27.0
XLM-R (Full) w/o ctx	45.4	39.2	47.6	30.2	34.8	34.3	30.2	33.0	33.6	33.4	28.5	30.9	37.5	30.0	32.4	36.7	32.1	31.7	30.4	32.8

Table 4: Overall per-language evaluation. The first three columns show the results on ARC Easy (E), ARC Challenge (C), and Regents 12 LivEnv (en). The following columns show the per-language and the overall results (the last column All) for all languages. *All* is the score averaged over all $E\chi\alpha\mu s$ questions.

Multilingual BERT (Devlin et al., 2019) is a fundamental multilingual model trained on 104 languages with a vocabulary of 110K word-pieces, with a total of 172M parameters (12 layers, 768 hidden states, 12 heads).

XLM-RoBERTa (Conneau et al., 2020) is a recent multilingual model based on RoBERTa (Liu et al., 2019c). It is trained on 100 languages, with a larger vocabulary of 250K sentence pieces. It comes in two sizes: *XLM-*R_{Base} (270M parameters, same architecture as mBERT, except vocab size), and *XLM-R* (550M parameters, 24 layers, 1,024 hidden states, 16 heads). For completeness, we include both in our experiments.

We fine-tuned the aforementioned models following the standard procedure for multiple-choice comprehension tasks, as described in (Devlin et al., 2019) and (Liu et al., 2019c), using the Transformers library (Wolf et al., 2019). The training details can be found in Appendix A.

5 Experiments and Results

In this section, we evaluate the performance of the baseline models described in Section 4 on the $E\chi\alpha\mu s$ dataset. In Table 4, we show the overall per-language performance of the evaluated models. The first group shows simple baselines: random guessing and IR over Wikipedia articles. IR is better than random guessing, but it is clear that most questions require reasoning beyond simple word matching. In the last group, we evaluate the knowledge contained in the models before and after the QA fine-tuning. First, we evaluate XLM-R as a knowledge base, and then we use the *Full* model but with the question–option pair only.

5.1 Multilingual Evaluation

The next two groups show (*i*) how continuous fine-tuning of XLM-R on multi-choice machine reading comprehension and multi-choice science QA helps, and (*ii*) how the different models (XLM-R, XLM-R_{Base}, and mBERT) compare. We follow a standard training scheme for such tasks: first we fine-tune on RACE (Lai et al., 2017) (∼85k EN questions over documents), then on the AI2 English science datasets (we call them SciENs for shorter), including ∼9k EN questions with provided relevant contexts[8], and, finally, on our multilingual training set (see Section 3.4) with retrieved relevant contexts from Wikipedia (see Appendix D), which is our desired multilingual evaluation setting and we call it *Full*. We can also see that training on the SciENs, which has mostly primary school questions from Natural Sciences, only yields +0.5% improvement on $E\chi\alpha\mu s$. Nevertheless, we see a 2.4% improvement with multilingual fine-tuning on $E\chi\alpha\mu s$ and +0.5% for English. In the third group, we compare the results from mBERT, XLM-R_{Base}, and XLM-R after fine-tuning. Increasing the capacity of the model yields improvements: XLM-R scores 7.4% higher on $E\chi\alpha\mu s$, and more than 14% on English datasets, compared to its base version (XLM-R_{Base}). However, mBERT and XLM-R_{Base} have close performance, with mBERT having a small advantage in the multilingual setting.

Finally, we fine-tuned mBERT on $E\chi\alpha\mu s$ only. As expected, the performance drops by 3% absolute compared to the *Full* setup.

[8]We use the data described at `http://leaderboard.allenai.org/arc/submission/blcotvl7rrltlue6bsv0`

Lang	A_E	A_{Ch}	R12	de	es	fr	it	pt	bg	hr	lt	mk	pl	sr	hu	sq	tr	vi	ar
en_{all}	73.6*	51.2*	68.4*	35.5*	37.9	37.1	39.5	39.3	44.2	38.5	41.3	49.8	36.1	37.4	37.9	42.5	37.4	35.9	39.1
w/ it	+1.4	+1.3	+1.4	+6.2	**+4.2***	+0.3*	-	-3.7*	+1.2	+4.1	+0.9	+0.8	+1.5	+3.1	+2.8	+0.9	-1.3	**+1.8**	+1.8
w/ pt	+0.1	+1.2	-0.8	+2.2	+2.5*	-2.5*	+1.4*	-	+0.3	0.0	+2.0	+0.8	-0.1	-0.6	-0.6	-1.3	**+1.3**	+0.6	+1.1
w/ bg	+0.6	+0.4	-0.4	**+3.6**	+0.8	+1.6	**+3.4**	-1.9	-	+1.5*	**+2.9***	**+1.6***	+0.1*	**+1.5***	+2.0	**+2.3**	-0.9	-0.8	+0.8
w/ hr	+1.1	**+1.7**	-0.2	+4.8	+3.8	+0.3	+5.8	-2.8	+1.7*	-	+0.2*	-0.1*	+1.2*	+6.7*	+2.8	+1.7	+1.2	+0.5	-0.1
w/ mk	+1.5	-0.5	**+2.2**	+1.0	**+4.2**	-0.3	+2.0	-2.6	+1.8*	**+3.9***	+1.5*	-	+1.9*	0.0*	+2.0	+6.9	+4.8	+0.5	**+4.5**
w/ pl	-2.0	-1.5	-3.1	0.0	+0.4	-2.5	+0.1	-1.3	+1.1*	+1.0*	-0.5*	-0.2*	-	0.0*	-0.4	+0.3	+0.2	-1.4	+0.9
w/ sr	**+1.8**	-0.1	-1.2	+2.6	+5.1	+1.9	+2.8	-0.6	**+2.2***	+6.2*	+0.2*	+1.3*	+1.3*	-	+1.4	-0.4	-0.7	-1.0	+3.2
w/ hu	-0.8	-0.8	-1.0	+7.8	+10.2	+2.8	+1.1	-1.9	+0.7	+0.8	-3.2	+0.1	+0.9	+0.9	-	-0.2	-0.2	-0.6	-1.4
w/ sq	-0.1	+0.3	-1.5	+3.5	-0.5	-0.6	+0.8	+0.9	+0.9	+0.8	+1.0	+3.4	+0.6	+0.6	+1.9	-	+0.4	+0.3	+0.2
w/ tr	-0.5	+1.1	-1.5	+1.5	+3.0	-1.9	+2.3	-3.0	+1.0	+1.0	-2.7	+1.5	+0.2	+1.2	**+2.4**	+3.7	-	-1.0	+1.8
w/ vi	-0.5	+0.4	-0.8	+2.9	+3.4	**+4.1**	+1.1	**+1.1**	+1.5	+1.7	+0.4	+0.4	**+2.1**	0.0	+1.7	+0.8	+1.1	-	+3.4

Table 5: Cross-lingual zero-shot performance on $E\chi\alpha\mu s$. The first three columns show the performance on the test set of the AI2 science datasets (English), followed by per-language evaluation. The underlined values mark languages that have parallel data with the source language, and the ones with an asterisk* are from the same family.

5.2 Knowledge Evaluation

The last two rows of Table 4 evaluate the knowledge in the best model, namely XLM-R. With *XLM-R as KB* (see Section 4.1) we see small improvement over the random baseline: +5% ARC Easy, 2% on R12, and just +1% on $E\chi\alpha\mu s$ and ARC Challenge. Furthermore, we evaluate the knowledge contained in the model after the *Full* fine-tuning by excluding the relevant knowledge context (*ctx*). This is better than the *XLM-R as KB*, but it still achieves inferior overall results, which shows that the stored knowledge is not enough, and that we need to explicitly obtain additional knowledge from an external source.

5.3 Cross-lingual Evaluation

Table 5 shows the results from the cross-lingual zero-shot transfer compared to the English-only baseline en_{all}, from XLM-R fine-tuned on SciEN. The languages are ordered by family, and then alphabetically. We further fine-tune on a single source language and we test on all other languages using the splits described in Subsection 3.4. The results show that the additional fine-tuning on a single language is mostly positive. This is notable when fine-tuning on a language with similar linguistic characteristics to the target language, e.g., Balto-Slavic: bg-sr, hr-mk, pl-mk, sr-bg.

We also see gains when the source language contains more questions from largely represented and harder subjects. Examples of such are the experiments showing the positive effects of training on Vietnamese and Macedonian as source languages; they both contain such subjects: Biology, History, Chemistry, Physics, and Geography.

This is an indication that the knowledge from the same or from related subjects in a non-related language is preferred over knowledge from non-related subjects from a related language. For the same reasons, Portuguese and Polish show negative effects of fine-tuning on some of the target languages. They contain mostly niche subjects such as Professional, Philosophy, Economics, Geology. We see a noticeable drop in accuracy for Portuguese almost everywhere, but it has positive effect on languages that contain similar subjects (Biology, Economics) or are from the same language family such as Spanish and Italian (for Portuguese). We see the opposite in the Lithuanian-Polish pair, languages from the same family (but different subjects) have negative, or no effect on each other. Finally, we analyze the results from language pairs containing parallel examples (the underlined values). Such pairs show consistent improvement (+5 to +10), which suggests that the model learns to align the parallel knowledge from the source language to the target language. However, we also must note that the effect is strongly dependent on the size of the overlapping sets.

5.4 Per-subject Fine-grained Evaluation

Fine-grained evaluation (Mihaylov and Frank, 2019; Xu et al., 2020) allows an in-depth analysis of the question answering models. One of the nice features of $E\chi\alpha\mu s$ is that it supports subject-related fine-grained evaluation. On Figure 5, the results are shown by subject group and per-subject for Natural Science.[9]

[9]Per-subject results for Social Science and Other are available in Appendix E.

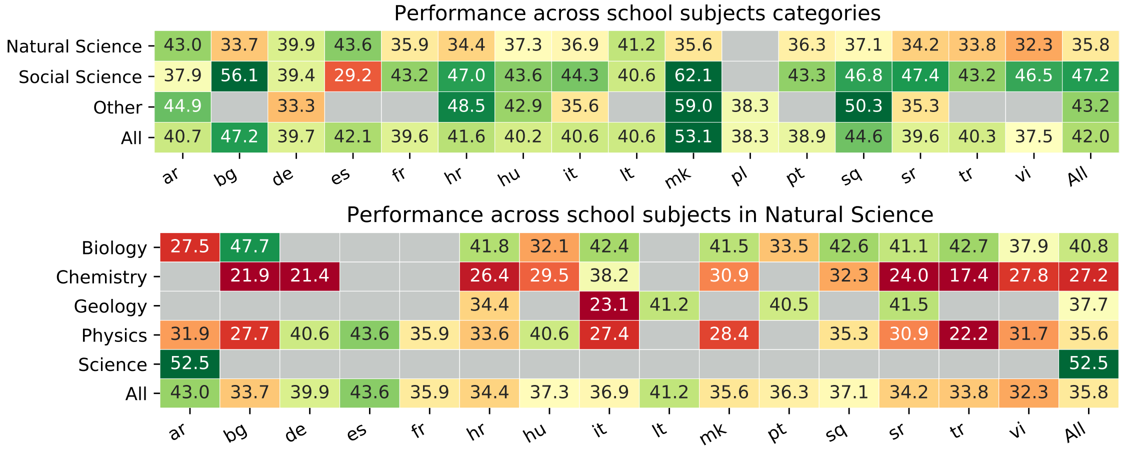

Figure 5: Fine-grained evaluation by language and school subjects.

We can see that the Natural Science questions are the most challenging ones, which is mostly due to Chemistry and Physics. Those questions require very complex reasoning and knowledge such as understanding physical models, processes and causes, comparisons, algebraic skills and multi-hop reasoning (see Section 3.5). These skills are currently beyond the capabilities of the current QA models, and pose interesting challenges for future work (Welbl et al., 2018; Yang et al., 2018; Saxton et al., 2019; Lample and Charton, 2020). Informatics is another challenging subject, as it requires understanding programming code and positional numerical systems among others.

6 Discussion

Our results show that initial fine-tuning on a large monolingual out-of-domain multi-choice machine reading comprehension dataset (RACE (Lai et al., 2017)) performs much better than *no training* baselines for answering multilingual $E\chi\alpha\mu s$ questions. Moreover, additional training on English science QA in lower school levels has no significant effect on the overall accuracy. These results suggest that further investigation of fine-tuning with other multilingual datasets (Gupta et al., 2018; Lewis et al., 2020; Clark et al., 2020; Efimov et al., 2020; d'Hoffschmidt et al., 2020; Artetxe et al., 2020; Longpre et al., 2020) is needed in order to understand the domain transfer benefits to science QA in $E\chi\alpha\mu s$, even if they are not in a multi-choice setting (Khashabi et al., 2020). Using *domain-adaptive* and *task-adaptive pre-training* (Gururangan et al., 2020) to the multilingual science QA might offer further potential benefits.

Moreover, we need a better knowledge context for a given question–choice pair (the last row in Table 4). Knowing that the context retrieved from the noisy Wikipedia corpus is relevant for answering $E\chi\alpha\mu s$ questions, suggests that we need a better multilingual science corpus, similar to Clark et al. (2018); Pan et al. (2019); Bhakthavatsalam et al. (2020). We further need better multilingual knowledge selection and ranking (Banerjee et al., 2019). Finally, our cross-lingual experiments show that we can align the knowledge between languages from parallel examples, which poses a new question: *Is it only due to keyword matching or could the model align full sentences?*

7 Conclusion and Future Work

We presented $E\chi\alpha\mu s$, a new challenging cross-lingual and multilingual benchmark for science QA in 16 languages and 24 subjects from high school examinations.

We further proposed new fine-grained evaluation that allows precise comparison across different languages and school subjects. We performed various experiments and analysis with pre-trained multilingual models (XLM-R, mBERT), and we demonstrated that there is a need for better reasoning and knowledge transfer in order to solve some of the questions from $E\chi\alpha\mu s$. We hope that our publicly available data and code will enable work on multilingual models that can reason about question answering in the challenging science domain.

In future work, we plan to extend the dataset with more questions, more subjects, and more languages. We further plan to develop new models to address the specific challenges we identified.

Acknowledgments

We thank the anonymous reviewers for their helpful questions and comments. We thank the AI2 Aristo Team for providing the data splits used for pre-training on SciEN datasets.

This research is partially supported by Project UNITe BG05M2OP001-1.001-0004 funded by the OP "Science and Education for Smart Growth" and co-funded by the EU through the ESI Funds.

References

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*, ACL '20, pages 4623–4637.

Pratyay Banerjee, Kuntal Kumar Pal, Arindam Mitra, and Chitta Baral. 2019. Careful selection of knowledge to solve open book question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL '19, pages 6120–6129, Florence, Italy.

Sumithra Bhakthavatsalam, Chloe Anastasiades, and Peter Clark. 2020. GenericsKB: A knowledge base of generic statements. *ArXiv*, abs/2005.00660.

Michael Boratko, Harshit Padigela, Divyendra Mikkilineni, Pritish Yuvraj, Rajarshi Das, Andrew McCallum, Maria Chang, Achille Fokoue, Pavan Kapanipathi, Nicholas Mattei, Ryan Musa, Kartik Talamadupula, and Michael Witbrock. 2018. An interface for annotating science questions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, EMNLP '18, pages 102–107, Brussels, Belgium.

Casimiro Pio Carrino, Marta R. Costa-jussà, and José A. R. Fonollosa. 2020. Automatic Spanish translation of SQuAD dataset for multi-lingual question answering. In *Proceedings of the 12th Language Resources and Evaluation Conference*, LREC '20, pages 5515–5523, Marseille, France.

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.

Peter Clark. 2015. Elementary school science and math tests as a driver for AI: Take the Aristo challenge! In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence*, AAAI '15, pages 4019–4021, Austin, Texas, USA.

Peter Clark, Niranjan Balasubramanian, Sumithra Bhakthavatsalam, Kevin Humphreys, J. Clint Kinkead, and Ashish Sabharwal. 2014. Automatic construction of inference-supporting knowledge bases. In *Proceedings of 4th Workshop on Automated Knowledge Base Construction*, AKBC '14, Montreal, Canada.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *ArXiv*, abs/1803.05457.

Peter Clark, Oren Etzioni, Daniel Khashabi, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, Niket Tandon, Sumithra Bhakthavatsalam, Dirk Groeneveld, Michal Guerquin, and Michael Schmitz. 2019. From 'F' to 'A' on the N.Y. regents science exams: An overview of the Aristo project. *ArXiv*, abs/1909.01958.

Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Turney, and Daniel Khashabi. 2016. Combining retrieval, statistics, and inference to answer elementary science questions. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI '16, page 2580–2586, Phoenix, Arizona, USA.

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*, ACL '20, pages 8440–8451.

Bhavana Dalvi, Niket Tandon, Antoine Bosselut, Wen-tau Yih, and Peter Clark. 2019. Everything happens for a reason: Discovering the purpose of actions in procedural text. 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 '19, pages 4496–4505, Hong Kong, China.

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*, NAACL-HLT '19, pages 4171–4186, Minneapolis, Minnesota, USA.

Martin d'Hoffschmidt, Maxime Vidal, Wacim Belblidia, and Tom Brendlé. 2020. FQuAD: French question answering dataset. *ArXiv*, abs/2002.06071.

Pavel Efimov, Andrey Chertok, Leonid Boytsov, and Pavel Braslavski. 2020. SberQuAD – Russian reading comprehension dataset: Description and analysis. In *Proceedings of the 11th International Conference of the CLEF Association: Experimental IR Meets Multilinguality, Multimodality, and Interaction*, CLEF '20, pages 3–15, Thessaloniki, Greece.

Deepak Gupta, Surabhi Kumari, Asif Ekbal, and Pushpak Bhattacharyya. 2018. MMQA: A multi-domain multi-lingual question-answering framework for English and Hindi. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, LREC '18, pages 2777–2784, Miyazaki, Japan.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 8342–8360.

Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2019. Beyond English-only reading comprehension: Experiments in zero-shot multilingual transfer for Bulgarian. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP '19, pages 447–459, Varna, Bulgaria.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of Machine Learning Research*, ICML '20, Online.

Yimin Jing, Deyi Xiong, and Zhen Yan. 2019. BiPaR: A bilingual parallel dataset for multilingual and cross-lingual reading comprehension on novels. 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 '19, pages 2452–2462, Hong Kong, China.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Peter Clark, Oren Etzioni, and Dan Roth. 2016. Question answering via integer programming over semi-structured knowledge. In *Proceedings of the Twenty-fifth International Joint Conferences on Artificial Intelligence Organization*, IJCAI '16, pages 1145–1152, New York, New York.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2018. Question answering as global reasoning over semantic abstractions. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, AAAI '18, pages 1905–1914, New Orleans, Louisiana, USA.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UnifiedQA: Crossing format boundaries with a single QA system. *Finding of EMNLP*.

Tushar Khot, Peter Clark, Michal Guerquin, Paul Edward Jansen, and Ashish Sabharwal. 2020. QASC: A dataset for question answering via sentence composition. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, AAAI '20, pages 8082–8090, New York, New York, USA.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2017. Answering complex questions using open information extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, ACL '17, pages 311–316, Vancouver, Canada.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A textual entailment dataset from science question answering. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, AAAI '18, pages 5189–5197, New Orleans, Louisiana, USA.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP '17, pages 785–794, Copenhagen, Denmark.

Guillaume Lample and François Charton. 2020. Deep learning for symbolic mathematics. In *Proceedings of the 8th International Conference on Learning Representations*, ICLR '20.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedings of the 8th International Conference on Learning Representations*, ICLR '20.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 7315–7330.

Seungyoung Lim, Myungji Kim, and Jooyoul Lee. 2019. KorQuAD1.0: Korean QA dataset for machine reading comprehension. *ArXiv*, abs/1909.07005.

Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2019a. XQA: A cross-lingual open-domain question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL '19, pages 2358–2368, Florence, Italy.

Pengyuan Liu, Yuning Deng, Chenghao Zhu, and Han Hu. 2019b. XCMRC: Evaluating cross-lingual machine reading comprehension. In *Proceedings of the International Conference on Natural Language Processing and Chinese Computing*, NLPCC '19, pages 552–564, Dunhuang, China.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019c. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv*, abs/1907.11692.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2020. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *ArXiv*, abs/2007.15207.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pages 2381–2391, Brussels, Belgium.

Todor Mihaylov and Anette Frank. 2019. Discourse-aware semantic self-attention for narrative reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, EMNLP '19, pages 2541–2552, Hong Kong, China.

Arindam Mitra, Peter Clark, Oyvind Tafjord, and Chitta Baral. 2019. Declarative question answering over knowledge bases containing natural language text with answer set programming. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, AAAI '19, pages 3003–3010, Honolulu, Hawaii, USA.

Jianmo Ni, Chenguang Zhu, Weizhu Chen, and Julian McAuley. 2019. Learning to attend on essential terms: An enhanced retriever-reader model for open-domain question answering. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, pages 335–344, Minneapolis, Minnesota, USA.

Xiaoman Pan, Kai Sun, Dian Yu, Jianshu Chen, Heng Ji, Claire Cardie, and Dong Yu. 2019. Improving question answering with external knowledge. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, MRQA '19, pages 27–37, Hong Kong, China.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '18, pages 2227–2237, New Orleans, Louisiana, USA.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, EMNLP-IJCNLP '19, pages 2463–2473, Hong Kong, China.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, EMNLP '16, pages 2383–2392, Austin, Texas, USA.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *ArXiv*, abs/2002.12327.

David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. 2019. Analysing mathematical reasoning abilities of neural models. In *Proceedings of the 7th International Conference on Learning Representations*, ICLR '19, New Orleans, Louisiana, USA.

Carissa Schoenick, Peter Clark, Oyvind Tafjord, Peter D. Turney, and Oren Etzioni. 2017. Moving beyond the Turing test with the Allen AI Science Challenge. *Communications of the ACM*, 60:60 – 64.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2019. Improving machine reading comprehension with general reading strategies. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, pages 2633–2643, Minneapolis, Minnesota, USA.

Oyvind Tafjord, Peter Clark, Matt Gardner, Wen tau Yih, and Ashish Sabharwal. 2019. QuaRel: A dataset and models for answering questions about qualitative relationships. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, AAAI '19, pages 7064–7071, Honolulu, Hawaii, USA.

Kiet Van Nguyena, Khiem Vinh Trana, Son T Luua, and Anh Gia-Tuan. 2020. Enhancing lexical-based approach with external knowledge for Vietnamese multiple-choice reading comprehension. *ArXiv*, abs/2001.05687.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, NIPS '17, pages 5998–6008, Long Beach, California, USA.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, W-NUT '17, pages 94–106, Copenhagen, Denmark.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Dongfang Xu, Peter Jansen, Jaycie Martin, Zhengnan Xie, Vikas Yadav, Harish Tayyar Madabushi, Oyvind Tafjord, and Peter Clark. 2020. Multiclass hierarchical question classification for multiple choice science exams. In *Proceedings of the 12th Language Resources and Evaluation Conference*, LREC '20, pages 5370–5382, Marseille, France.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32*, NIPS '19, pages 5753–5763. Vancouver, Canada.

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*, EMNLP '18, pages 2369–2380, Brussels, Belgium.

A Fine-Tuning and Hyper-parameters

In this work, we are interested in the cross-lingual transferability of multilingual models such as mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020), each of which comes pre-trained on more than 100 languages. We evaluated the QA capabilities of these models, following the established protocol (Devlin et al., 2019; Liu et al., 2019c; Sun et al., 2019), namely we fine-tuned them to predict the correct answer in a multi-choice setting, given a selected context. The aforementioned setup feeds the pre-trained model with a text, processed using the model's tokenizer in the following format:

[CLS] C [SEP] Q + O [SEP]

where C, Q and O are the tokenized *knowledge **C**ontext* (see Appendix D), ***Q**uestion*, and ***O**ption*, respectively.

We used the Transformers library (Wolf et al., 2019). We fine-tuned mBERT, XLM-R, and XLM-R_{Base} in three steps. We first fine-tuned the models with RACE (Lai et al., 2017), a multiple-choice reading comprehension dataset with around 85k questions for training. Then, we trained on the combination of ARC (Clark et al., 2018), OpenBookQA (Mihaylov et al., 2018), and Regents Living Environments, as in the *AristoRoBERTaV7* ARC Challenge leaderboard entry[10]; we refer to these datasets as *SciENs* (**Sci**ence **E**nglish dataset**s**). We used the resulting pre-trained models as base models for our *Multilingual* and *Cross-lingual* evaluations (Section 5 in the paper). For the multilingual evaluation, we continued training the model, previously fine-tuned on the SciENs datasets, with our multilingual Train$_{Mul}$ set, validating on Dev$_{Mul}$ and testing on Test$_{Mul}$. For our cross-lingual evaluation, we continued training the SciENs model on separate languages, as described in Section 5.3.

In Table 6, we show the values of the hyper-parameters for each fine-tuning step and corresponding model. Note that these hyper-parameters were **not** obtained with an exhaustive search, and thus a better setting might exist for each model and dataset. Initially, we used the hyper-parameters for *AristoRoBERTaV7* ARC Leaderboard submission for English-only RoBERTa (Liu et al., 2019c): epochs = 4, learning rate = 1e-5.

[10]https://leaderboard.allenai.org/arc/submission/blcotvl7rrltlue6bsv0

With these parameters alone, the models did not perform well, and thus we added a warmup of 0.1 and a weight decay of 0.06, which stabilized the training. In all experiments, we used the Adam optimizer with β_1=0.9, β_2=0.999, and ϵ=1e-08.

We further performed manual tuning of the hyper-parameters: we experimented with variations thereof, depending on the performance on the corresponding development sets, and we ended up with the values in Table 6. Moreover, we adjusted the batch size and the accumulation steps depending on the availability of the GPUs on our cluster: Nvidia GTX 1080 Ti (Pascal, 11GB memory) or Nvidia Quadro RTX 6000 (24GB). For each examined setting, we trained for up-to 6 epochs, evaluating the model on the corresponding development set every 100 to 1000 update steps, depending on the dataset size and the effective batch size. For the final evaluations, we chose the model with the highest accuracy score on the corresponding development set.

Fine-tuning XLM-R (550M parameters) on Nvidia Quadro RTX 6000 (24GB) with the given hyper-parameters took around three hours per epoch when fine-tuned on RACE ($\sim$85k examples), 30 minutes per epoch when fine-tuned on SciENs ($\sim$9k examples), and 30 minutes on $E\chi\alpha\mu s$ on Train$_{Mul}$ ($\sim$8k examples). Fine-tuning XLM-R_{Base} (270M parameters) and mBERT (172M parameters) on Nvidia GTX 1080 Ti (Pascal, 11GB memory) with the given hyper-parameters took roughly 2 to 2.5 hours per epoch when fine-tuned on RACE ($\sim$85k examples), 30 to 35 minutes per epoch when fine-tuned on SciENs ($\sim$9k examples), and additional 30 minutes on the $E\chi\alpha\mu s$ Train$_{Mul}$ ($\sim$8k examples).

B Subject Analysis

The *Natural Science* group contains five subjects. The corresponding question length is 16.4 characters and 3.9 answers on average. Some of the subjects are well-known and widely studied, such as *Physics, Biology* and *Chemistry*. They appear in at least 10 out of the 16 languages, covering 7 out of 8 language families. However, *Geology* is less common and is present for only 4 languages. Finally, *Science* is an isolated subject for Arabic. This group contributes a total of 9,962 questions in the entire dataset, as shown in Table 7. The major groups in the table are divided with a horizontal line for convenience.

Model	Batch Size	Accum. Steps	Max Seq. Len.	Learn Rate	Warmup	Weight Decay
			fine-tune on RACE (Step 1)			
mBERT	4	64	320	0.00005	0.1	-
XLM-R XLM-R_{Base}	2	16	320	0.00001	0.1	0.06
			fine-tune on SciENs (Step 2)			
mBERT XLM-R XLM-R_{Base}	2	16	320	0.00001	0.2	0.06
			$E\chi\alpha\mu s$ Train$_{Mul}$ (Step 3 - Multilingual)			
mBERT XLM-R XLM-R_{Base}	2	16	320	0.00001	0.2	0.06
			for each source language (Step 3 - Cross-lingual)			
mBERT XLM-R XLM-R_{Base}	2	8	320	0.00001	0.2	0.06

Table 6: The hyper-parameter values we used for fine-tuning.

The second subject group covers *Social Sciences*. *Geography, History, Philosophy, Psychology* and *Ethics* are more common, and thus are included in seven languages on average (see Table 7). The subject group's average question length is 18.5 characters. The only sizeable deviation being for Citizenship, as most of the questions in this subject explain some social situation in detail.

The last and smallest of the three subject groups is *Others*. It combines subjects that cannot be categorized as exactly science-related (either social or natural). Those subjects are often specific for a particular country or culture and are fairly diverse. As expected, they are present for less languages (just two).

B.1 Subject Definitions

Next, we give a brief description of the less commonly known subjects included in our dataset.

Agriculture covers questions about soil farming and preservation, small animals breeding and their general health care, and vehicle maintenance and repair.

Business & Economics is a term used to combine five similar subjects related to business and economics. The questions in these subjects cover theoretical questions on economics basis, marketing questions, business questions with elements of accountancy, finances, and organizational studies.

Citizenship is a specific subject from the Vietnamese school system, which tries to inform and give better perspective on different social situations, to educate students in how to perform better, and to be a more aware member of the society by analyzing different norms and personal morality.

Fine Arts contains analytical and historical questions about different forms of art such as movies, music, art, etc.

Forestry studies the craft of managing, using, conserving, and repairing forests, woodlands, and associated resources around them such as water sources and soil.

Geology is the study of the Earth, with the general exclusion of present-day life, flow within the ocean, and the atmosphere. Questions from this subject cover branches of Geology such as Economical Geology, Marine Geology, Geomorphology, and Geophysics.

Informatics consists of questions about basic hardware knowledge and software management as well as basics of different positional numeral systems (e.g., binary and hexadecimal).

Islamic Studies refers to the academic studies of Islam, Quran excerpts, and Muslim morality. This a subject studied in the Qatari educational system during both middle and high school.

Group	Subject	Language	Grade	Q Len	Ch Len	#Ch	#Q	Vocab
Natural Science	Biology	ar, bg, hr, hu, it, sr, sq, mk, tr, pt, vi	H	18.2	4.6	4.0	3,042	24,603
Natural Science	Chemistry	bg, hr, it, sr, de, hu, sq, mk, tr, vi	H	17.3	4.6	4.2	2,315	14,420
Natural Science	Geology	hr, it, sr, lt, pt	H	12.9	5.6	4.0	720	7,251
Natural Science	Physics	ar, bg, hr, it, sr, fr, de, hu, es, sq, mk, tr, vi	H	24.9	7.0	3.6	3,465	26,103
Natural Science	Science	ar	M, H	9.1	3.0	4.0	120	1,239
Social Science	Busin. & Econ.	fr, de, hu, sq, mk, tr, pt	H	5.7	6.5	3.9	2,012	16,875
Social Science	Citizenship	vi	H	45.1	6.3	4.0	119	980
Social Science	Ethics	hr, it, sr	H	15.5	2.6	4.0	194	1,859
Social Science	Geography	bg, hr, fr, de, hu, it, sr, es, tr, vi	H	15.2	5.0	4.2	1,349	11,207
Social Science	History	bg, hr, it, sr, lt, sq, mk, tr, vi	H	16.6	5.9	4.1	3,300	32,709
Social Science	Philosophy	bg, hr, it, sr, sq, mk, tr, pt	H	16.5	3.9	4.1	1,903	19,373
Social Science	Politics	hr, hu, it, sr	H	18.2	2.8	3.0	493	5,068
Social Science	Psychology	hr, it, sr	H	16.5	3.9	4.1	1,903	19,373
Social Science	Social	ar	M, H	10.8	3.4	4.0	277	2,828
Social Science	Sociology	hr, it, sr, sq, mk, tr	H	15.2	3.4	4.0	566	6,374
Other	Agriculture	hu	H	7.9	3.6	4.3	215	1,918
Other	Fine Arts	sq, mk	H	12.1	3.8	4.0	757	5,691
Other	Forestry	hu	H	7.8	2.9	3.7	241	1,957
Other	Informatics	hr, it, sr	H	18.7	6.2	4.0	311	2,695
Other	Islamic Studies	ar	M, H	9.4	3.0	4.0	78	925
Other	Landscaping	hu	H	7.4	3.8	4.9	49	596
Other	Professional	pl	H	13.7	4.3	4.0	1,971	18,990
Other	Religion	hr, sr	H	10.3	3.6	4.0	222	2,159
Other	Tourism	de, hu	H	8.8	5.2	4.0	20	359

Table 7: Per-subject statistics. The grade is High (H), and Middle (M). The average length of the question (*Q Len*) and the choices (*Ch Len*) are measured in number of tokens, and the vocabulary size (*Vocab*) is shown in number of words.

Landscaping teaches about modifying the visible features of an area of land, trees and park decorations. It also contains questions about plants and soils.

Politics covers Croatia's political system, historical questions about the country's development, as well as different regulations and laws, international relations and contracts.

Professional subject is present in the Polish school system and covers knowledge on specific professions such as flight attendant, babysitter, care taker, office worker in terms of profession's regulations, rules and established norms, etc.

Religion subject covers Christianity studies such as Bible knowledge, related traditions, e.g., baptism, marriage, etc.

Tourism covers hospitality management, as well as basis of business and traditions in Hungary and its neighbouring countries.

Science which is used in the Arabic school system throughout middle and high grade studies combines general science questions from Biology, Chemistry, Physics Geology and their branches such as as Biophysics, Astrophysics, and Biochemistry.

Social subject, similarly to Science, combines questions from political, cultural, historical and geographical studies.

Sociology is the study of society, patterns of social relationships, social interaction, and culture that surrounds our everyday life.

Language	Wiki code	#Sentences (millions)	#Articles (millions)	Stop word removal	Stemming	Keyword extraction	Language specific
ARC Corpus	-	14.6	-	✓	✓	✓	✓
German	de	50.0	2.43	✓	✓	✓	✓
French	fr	30.0	2.22	✓	✓	✓	✓
Italian	it	17.5	1.61	✓	✓	✓	✓
Spanish	es	22.7	1.60	✓	✓	✓	
Polish	pl	15.6	1.41		✓		
Vietnamese	vi	6.4	1.25	✓	✓		✓
Portuguese	pt	11.6	1.03	✓	✓	✓	
Arabic	ar	6.0	1.04	✓	✓	✓	✓
Serbian	sr	4.6	0.63				
Hungarian	hu	7.1	0.47	✓	✓	✓	
Turkish	tr	4.0	0.35	✓	✓	✓	✓
Bulgarian	bg	3.0	0.26	✓	✓	✓	
Croatian	hr	2.7	0.22				
Lithuanian	lt	2.0	0.20	✓	✓	✓	
Macedonian	mk	1.6	0.11				
Albanian	sq	0.8	0.08				

Table 8: Description of the per-language indices used as a source of background knowledge in our experiments.

C Reasoning and Knowledge Types

For our reasoning and knowledge type annotations, we followed the procedure and re-used the annotation types presented in (Clark et al., 2018; Boratko et al., 2018). However, as they were designed mainly for Natural Science questions, we had to extend them with two new types:

Domain Facts and Knowledge (Knowledge) This skill requires specific expertise in properties and facts in a given domain, e.g., physical properties, characteristics of a chemical element.

Example from Philosophy (*Portugal*):

Which of the following is an example of a priori knowledge?
A) I know my name.
B) I know how old I am.
C) *I know that no brother is an only child.* ✓
D) I know some parents are not married.

Negation (Reasoning) is a direct statement of negation, and it is often combined with other reasoning types such as linguistic matching.

Example from Fine Arts (*North Macedonia*):

Which of the following works of art does ***not*** *belong to the fine arts?*
A) Graphics.
B) *Poem.* ✓
C) Design.
D) Sculpture.

D Background Knowledge Corpus

Students need good textbooks to study before they can pass an exam, and the same holds for a good machine reading model. However, finding the information needed to answer a question, especially for questions in such a narrow domain as the subjects studied in high schools, usually requires a collection of specialized texts. The ARC Corpus (Clark et al., 2018) is an example of such a collection. It is built by querying a major search engine, and around 100 hand-written templates for 80 science topics covered by US elementary and middle schools. Albeit effective, this strategy relies on crafting templates for all language–subject pairs, making the task time-consuming if applied to multiple languages and subjects.

In our work, we used articles from Wikipedia to build a background knowledge corpus for each language. In particular, we parsed the text from the entire Wikipage, removing non-textual content, e.g., HTML tags, tables, etc. Following the common strategy used to solve similar tasks in English (Clark et al., 2018; Mihaylov et al., 2018), we split each document into sentences and we indexed them using an inverted index. In order to reduce the search space, and to mitigate the effect of known linguistic phenomena within the same language family, e.g., homonyms, partially shared alphabet, etc., we created a separate index for each language.

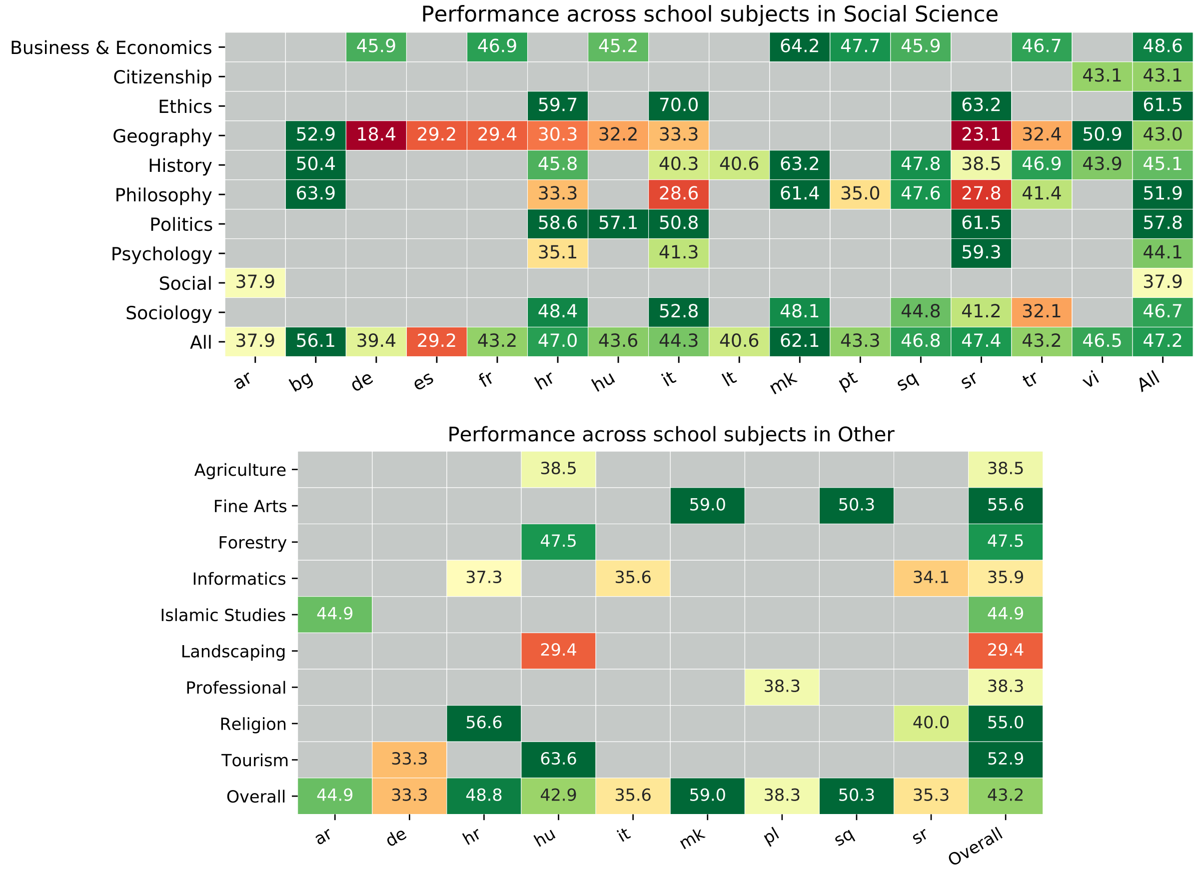

Figure 6: Fine-grained evaluation by language and school subjects in *Social Science* and *Other*.

Table 8 describes the main characteristics of the indices created for each language from its Wikipedia dump.[11] We compared the size of our index to the one from ARC (Clark et al., 2018). The number of articles for each language is taken from Wikipedia's official statistics [12]. We also marked the language analysis applied on the index. Some of the languages in $E\chi\alpha\mu s$ are low-resource ones, especially the ones from the Balto-Slavic family, which is also clear from their Wikipedia sizes. In the table, we see that half of the languages have under one million articles, and Albanian even falls under 100K. Moreover, even more languages are comparable with the number of sentences in the ARC Corpus, which is also built from science books. Finally, some of the languages (Serbian, Croatian, Macedonian, and Albanian) are not processed with any language-specific ElasticSearch analyzers.

E Fine-Grained Evaluation

Figure 6 shows fine-grained evaluation for two subject groups: *Social Science* and *Others*. We can see that these subjects are less challenging than Natural Science. One reason is that many of the subjects in these two groups such as Business & Economics, Geography, and History can be answered using knowledge that is easily accessible in sources such as Wikipedia (e.g., "*Who was the first prime minister of Poland after 1990?*"), i.e., without the need for complex reasoning or calculations, which are often needed in order to answer questions in subjects such as Physics and Chemistry. Nevertheless, while seeing scores as high as 60% for some subjects and languages, the current multilingual QA models are still far from perfect, which leaves a lot of room for improvement.

[11] We used the official Wiki dumps from March 2020 for all languages. http://dumps.wikimedia.org/

[12] The statistics are extracted from http://meta.wikimedia.org/wiki/List_of_Wikipedias