M2QA: Multi-domain Multilingual Question Answering

Anonymous ACL submission

Abstract

Generalization and robustness to input variation are core desiderata of machine learning research. Language varies along several axes, most importantly, language instance (e.g. French) and domain (e.g. news). While adapting NLP models to new languages within a single domain, or to new domains within a single language, is widely studied, research in joint adaptation is hampered by the lack of evaluation datasets. This prevents the transfer of NLP systems from well-resourced languages and domains to non-dominant language-domain combinations. To address this gap, we introduce M2QA, a multi-domain multilingual question answering benchmark. M2QA includes 13,500 SQuAD 2.0-style question-answer instances in German, Turkish, and Chinese for the domains of product reviews, news, and creative writing. We use M2QA to explore cross-lingual cross-domain performance of fine-tuned models and state-of-the-art LLMs and investigate modular approaches to domain and language adaptation. We witness **1)** considerable performance *variations* across domain-language combinations within model classes and **2)** considerable performance *drops* between source and target language-domain combinations across all model sizes. We demonstrate that M2QA is far from solved, and new methods to effectively transfer both linguistic and domain-specific information are necessary.[1]

1 Introduction

One of the central goals of natural language processing (NLP) is to develop systems that generalize well across different distributions, such as texts in different languages and domains.[2] While Transformer models have brought tremendous progress in NLP in recent years, especially evident with the recent emergence of large language models (LLMs), the problem of generalizing to new domains and languages is still far from solved. In-context learning (ICL), which refers to the ability of LLMs to perform tasks based on examples or instructions in the input prompt (Brown et al., 2020), is likely the reason for their emergent abilities (Lu et al., 2023) – yet, even with in-context learning, Transformers cannot generalize beyond their pre-training data (Yadlowsky et al., 2023), and their performance varies considerably across languages and is particularly low in languages underrepresented in the training data (Laskar et al., 2023).

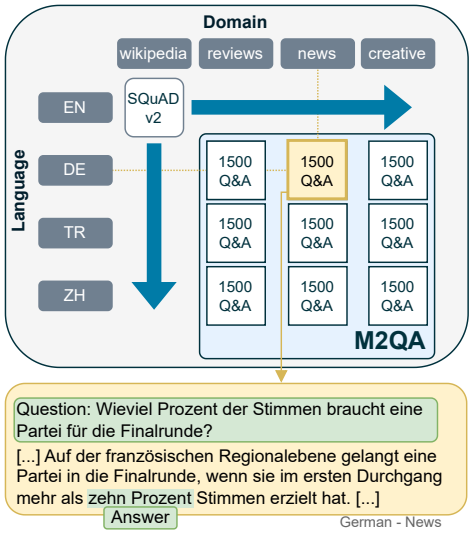


Figure 1: M2QA enables joint multi-domain and multilingual QA evaluation of NLP models across three diverse languages and three distinct domains (top) with 1500 SQuAD 2.0-style question-answer pairs for each language-domain combination (bottom).

With over 7,000 documented languages (Joshi et al., 2020) and countless domains, ensuring sufficient pre-training data coverage for every possible language-domain pair is hardly feasible. This motivates the development of methods that allow NLP systems to adapt to new languages and domains. While isolated language (e.g. Conneau et al., 2018; Hu et al., 2020; Artetxe et al., 2020b; Ghaddar

[1] We make M2QA publicly available upon acceptance

[2] Domains defined as text associated with a specific topic, such as product reviews or news (Gururangan et al., 2020).

and Langlais, 2017; Scialom et al., 2020) and domain adaptation (e.g. Wang et al., 2018) are extensively covered in prior work, the lack of comprehensive multi-domain multilingual benchmarks makes it difficult to objectively evaluate joint language and domain transfer methods. Existing multi-domain multilingual benchmarks either contain only one language in addition to English (Gupta et al., 2018), use machine-generated text (Bassignana et al., 2023) or are task-oriented dialogue systems that use narrow application-specific domains rather than a diverse set of domains useful for a wide range of applications (Moghe et al., 2023; Hu et al., 2023a). Results on these benchmarks suggest that language and domain are not independent axes. Therefore, we cannot infer the performance of joint transfer from individual axes, making it hard to systematically compare NLP models across languages and domains and to study joint language and domain adaptation approaches.

To address this gap, we introduce M2QA, a multi-domain multilingual SQuAD 2.0-style (Rajpurkar et al., 2016) extractive question answering (QA) dataset. We manually annotate naturally occurring texts in the respective languages – as opposed to translating documents from English – in order to increase lexical diversity (Rabinovich et al., 2016), mitigate the introduction of artifacts (Artetxe et al., 2020a) such as "translationese" (Bizzoni et al., 2020), and integrate the cultural idiosyncrasies of the target language (Hershcovich et al., 2022; Kuulmets and Fishel, 2023). The new benchmark makes it possible to study how well existing models perform at joint language and domain transfer (RQ1), whether specific language-domain combinations are especially hard to tackle for current models (RQ2), and whether existing methods (e.g. full fine-tuning, modular setups, ICL / instruction-based methods for LLMs) compare on language and domain transfer (RQ3).

In summary, our paper makes the following contributions: **1)** We create a multi-domain multilingual extractive QA benchmark, covering three domains (product reviews, news, creative writing) and three languages (German, Turkish, Chinese), resulting in 13,500 answerable and unanswerable QA instances (Figure 1). **2)** We evaluate baseline and transfer performance using a wide range of models and transfer techniques, including fully-finetuned models, modular transfer learning and LLMs. **3)** We find that transfer performance considerably varies across domain-language combinations. **4)** We find that the widely used SQuAD 2.0 evaluation metric implementation is insufficient for evaluating multilingual extractive QA due to its reliance upon whitespace tokenization and propose a version of the metric that mitigates the issue. **5)** Our results show that modern LLMs perform considerably worse on their target than on their source domain-language pair, highlighting the need for further research into methods that transfer both linguistic *and* domain-specific information.

2 Background

2.1 Adaptation and Modularity

Transfer and adaptation methods aim to optimize model performance on unseen data distributions. This can be achieved through modular deep learning methods (Pfeiffer et al., 2023) that combine modules containing knowledge about different aspects of the task, such as the language or domain. Modular approaches may involve merging weights of individually trained models (Ilharco et al., 2023) or model ensembling (Blevins et al., 2024; Li et al., 2022). However, these methods require fully fine-tuning multiple models. Parameter-efficient fine-tuning (PEFT) or adapter methods (Houlsby et al., 2019; Hu et al., 2022; Ben Zaken et al., 2022; Ansell et al., 2022) overcome these limitations. Instead of updating all model weights in the fine-tuning stage, adapter methods only fine-tune a small set of parameters while keeping the majority of parameters frozen.

2.1.1 Domain and Language Transfer

Domain transfer is the process of learning a task on a set of domains and then applying the model to the same task in a previously unseen domain. Domain transfer can be accomplished by (sequentially) fine-tuning LMs on in-domain data (Howard and Ruder, 2018; Pruksachatkun et al., 2020; Poth et al., 2021; Gururangan et al., 2020) or by combining multiple expert LMs (Li et al., 2022), where new domains are added by training new expert LMs. Gururangan et al. (2023) propose to cluster the data in the beginning to avoid massive node synchronization. To avoid full model fine-tuning, domain-specific adapters have been used (Chronopoulou et al., 2022, 2023). Gururangan et al. (2022) replace the Transformer's feedforward layers with DEMIX layers consisting of multiple domain experts. In this modular solution, the DEMIX layers

of different domains can be combined to handle heterogeneous domains during inference.

For language transfer, a model trained on a task in one or more source language(s) is evaluated on a different target language. While LLMs such as GPT-4 (OpenAI, 2023) can perform zero-shot or few-shot language transfer between similar languages, smaller models – such as XLM-Roberta (Conneau et al., 2020) – need to be fine-tuned or otherwise adapted: Blevins et al. (2024) combine multiple expert LMs, similar to the domain branch-train-merge setup; modular approaches (e.g. Pfeiffer et al., 2020; Ansell et al., 2021; Parović et al., 2022; Parovic et al., 2023) train language-specific and task-specific adapters to perform language transfer by exchanging the language adapter. Modular setups also find application in the joint transfer between domain and language. Cooper Stickland et al. (2021) use domain and language-specific adapters to transfer to languages and domains. The m^4 adapter (Lai et al., 2022) uses meta-learning with adapters for multi-domain multilingual machine translation. Kulkarni et al. (2023) propose a mixture-of-experts to perform multi-domain multilingual named entity recognition.

In this paper, we explore variations of all previously mentioned adaptation techniques: **1)** fully fine-tuning smaller models; **2)** a modular setup following the MAD-X method; **3)** zero-shot and few-shot approaches using LLMs.

2.2 Evaluation

Domain and language transfer techniques are mainly evaluated based on perplexity (e.g. Li et al., 2022; Gururangan et al., 2022; Conneau and Lample, 2019) or downstream tasks (e.g. Pfeiffer et al., 2020; Gupta et al., 2023). Perplexity is a token-level metric which overemphasizes the importance of frequent tokens and constructions (Dudy and Bedrick, 2020) and does not necessarily account for task-specific phenomena. Hence, it is questionable if perplexity is a good indicator of downstream task performance.

Question answering has been used to evaluate language and domain transfer separately. Prominent multilingual datasets are XQuAD (Artetxe et al., 2020b) and MLQA (Lewis et al., 2020). For domain transfer, the Quail benchmark (Rogers et al., 2020) provides a multiple-choice QA dataset. MultiReQA (Guo et al., 2021) combines existing QA datasets to a new multi-domain benchmark.

Benchmarks that target cross-lingual and cross-domain transfer in other tasks than QA also exist; MultiFC (Augenstein et al., 2019) and CrossRE (Bassignana and Plank, 2022) contain multiple domains for the same task. M2D2 (Reid et al., 2022) introduces a massively multi-domain setup with 145 subdomains evaluating performance with perplexity. Chronopoulou et al. (2022) evaluate perplexity across domains found on websites. Other popular cross-lingual tasks are NER (e.g. Ghaddar and Langlais, 2017) and summarization (e.g. Scialom et al., 2020). Most NLP benchmarks only focus on exploring one dimension, i.e. multilinguality *or* multi-domain (Ruder et al., 2022), which prevents investigating non-linear dependencies between domain and language. We discuss this in more detail in Section 3.1 below.

3 M2QA Dataset

3.1 Requirements

We define the following requirements for a benchmark that allows joint evaluation of language and domain transfer methods: (R1) Coverage: The benchmark should provide annotated data for each language-domain combination. (R2) Diversity: The benchmark should cover typologically distinct language and a broad range of domains. (R3) Openness: The source texts should be open-licensed and available for research usage. (R4) Universal task: The data should be annotated using a domain-agnostic task, enabling cross-domain comparison.

An important trade-off pertains to the use of translated vs. naturally occurring texts. Translated texts ensure that the data covers the same topics within the domain, resulting in aligned text across languages. However, translations have lower lexical diversity (Rabinovich et al., 2016) and introduce artifacts (Artetxe et al., 2020a) such as unnatural language usage and "translationese" (Bizzoni et al., 2020). Hershcovich et al. (2022) show that culture affects several axes of text variation. Translations contain the cultural background of the source language that does not correspond to the cultural background of native speakers of the target language (Kuulmets and Fishel, 2023). We prioritize language representative of how native speakers write over aligned text. Thus, we require (R5) Naturalness: all texts in the benchmark should have been produced naturally, not via translation.

Few multilingual and multi-domain datasets have been previously proposed. MMQA (Gupta

et al., 2018) includes factoid and short descriptive questions in English and Hindi over 6 domains. Multi3WOZ (Hu et al., 2023b) and Multi3NLU++ (Moghe et al., 2023) are multi-domain and multilingual benchmarks for task-oriented dialogue. README++ (Naous et al., 2023) is a multi-domain multilingual benchmark for readability assessment which includes translated texts in some of the domains. CrossRE (Bassignana and Plank, 2022) is a machine-generated, human-verified, multi-domain, multilingual benchmark for relation extraction. As Table 1 shows, none of the existing datasets fulfil our requirements as defined above.

3.2 Design

As per our requirements, the languages and domains in M2QA should cover a variety of language families and text styles (R2) to ensure that the transfer is not trivial. We chose German (Indo-European Germanic), Turkish (Turkic), and Chinese (Sino-Tibetan) as languages. As domains, we chose product reviews, news, and creative writing, covering various writing styles, levels of formality, and vocabularies. To fulfil R1, we annotated data for every language and domain combination. We collected open (R3) texts that are originally written in the target language to ensure naturalness (R5).

The annotated task needs to be universal (R4). One universal task is extractive question answering (QA). For extractive QA, the input is a question and a context that provides information to answer the question. The task is to extract the shortest span from the context that answers the question or, if the context does not contain an answer to the question, return that the question is unanswerable. An example question is shown in Figure 1. Extractive QA requires natural language understanding to identify the information needed to answer the question. Additionally, it requires reasoning to connect the concepts mentioned in the question to those mentioned in the text and extract the span with the relevant information. This makes extractive QA a complex task suitable for our benchmark.

3.3 Dataset Creation

Our annotation process consists of three parts: Passage curation, annotation and quality assurance.

Passage Curation. Collecting a benchmark that contains multiple languages and domains is not trivial, as the language and domain are entangled. Additionally, the data size varies for different domain and language combinations. For instance, scientific texts are mostly written in English. During the creation of the M2QA benchmark, we collected task annotations from combinations that are non-trivial to find. For instance, with our requirement for the data to be open (R3), finding creative writing data is challenging as most books have a copyright. For product reviews, we use the Chinese and German parts of MARC (Keung et al., 2020) and the Turkish product reviews dataset.[3] For news, we use the German 10kGNAD (Schabus et al., 2017) dataset, the Chinese CNewSum (Wang et al., 2021), and Turkish BilCat (Toraman et al., 2011). The creative writing domain is covered by German books from the Gutenberg Corpus (Gerlach and Font-Clos, 2018) and Turkish and Chinese stories published on Wattpad[4] with an open license. For more details on the data sources, licensing information, and preprocessing, see Appendix A.

Annotation. For the question-answer collection, we hired crowdworkers from Prolific[5], which was chosen due to its high annotation quality (Douglas et al., 2023) and advanced annotator filtering options. For each passage, the crowdworkers provided three answerable and two unanswerable questions. For answerable questions, they selected the shortest text span of the passage that answers the question. Following SQuAD 2.0 (Rajpurkar et al., 2018), we also let crowdworkers select a plausible answer span for unanswerable questions to make them harder to classify. We limit the maximum answer length to fall within 97% of the answers in XQuAD: ten words for German, nine for Turkish, and 22 characters for Chinese. For details on the annotation process, see Appendix H.

Quality Assurance To promote high data quality, crowdworkers were required to have at least a Bachelor's degree, speak the language in which they annotate data as their first language, and be fluent in English to understand the tutorial[6]. After the first annotation session, we manually reviewed ten randomly sampled question-answer pairs for each annotator, including at least one answerable and one unanswerable question. We translated annotations with DeepL.[7] If more than one QA pair violated our guidelines, we excluded the annotator's data from the dataset and removed the annota-

[3] https://huggingface.co/datasets/turkish_product_reviews
[4] https://www.wattpad.com/
[5] https://www.prolific.com
[6] The tutorial can be found here: Link upon acceptance
[7] https://www.deepl.com/api

Dataset	Task	Coverage (R1)	Diversity (R2)	Openness (R3)	Universal Task (R4)	Naturalness (R5)
MMQA (Gupta et al., 2018)	QA	✓	✗	✓	✓	✓
Multi3WOZ (Hu et al., 2023b)	ToD	✓	✓	✓	✗	✓
Multi3NLU++ (Moghe et al., 2023)	ToD	✓	✗	✓	✗	✗
README++ (Naous et al., 2023)	RA	✗	✓	✓	✗	(✓)
CrossRE (Bassignana and Plank, 2022)	RE	✓	✓	✓	✓	✗
M2QA	QA	✓	✓	✓	✓	✓

Table 1: Overview of existing multilingual multi-domain datasets along with their key characteristics and task. (QA = Question Answering, ToD = Task-oriented Dialogue, RA = Readability Assessment, RE = Relation Extraction).

tor from the worker pool. The results of the manual checks can be found in Appendix G. In total, we employed 162 crowdworker annotators, of which 19% (31 annotators) were rejected for poor-quality questions. From the questions kept, we manually checked 1310 questions (9.7% of the dataset)

3.4 Statistics

We collected 1500 question-answer pairs for every domain-language combination, resulting in 13,500 question-answer pairs. The domains are lexically diverse: maximum Jaccard similarity between domains is 0.135 in German, 0.115 in Turkish and 0.169 in Chinese (Appendix A.1). The average answer length is 3.62 words in German, 3.06 in Turkish and 4.46 in Chinese, similar to XQuAD (Artetxe et al., 2020b) with 2.98 words in German, 2.92 in Turkish, and 3.51 in Chinese respectively.

4 Experiments

The curation of the M2QA benchmark allows us – for the first time – to explore the transfer capabilities of state-of-the-art LMs along multiple dimensions. We will use M2QA to investigate the following research questions: **(RQ1)** How well do existing models perform at transfer learning across language and domains jointly? **(RQ2)** What language domain combinations are especially hard to tackle for the current models? **(RQ3)** How do modular adapter-based methods compare to fully-finetuned models in domain and language transfer?

4.1 Base Models

We first introduce our baseline models. See Appendix B for details on XLM-R models; Appendix D.2 lists the LLM prompts.

XLM-R$^{\text{Base}}$ (Conneau et al., 2020) is a multilingual Transformer encoder based on RoBERTa (Liu et al., 2019) that has been extensively studied in prior research on adaptation. We fine-tune the model on the English Wikipedia SQuAD 2.0 dataset (Rajpurkar et al., 2018) and evaluate it on different languages and domains of the M2QA benchmark. For data samples from languages other than English and not from the Wikipedia domain, this requires transfer across both dimensions.

XLM-R$^{\text{Domain}}$ As a second baseline, we evaluate XLM-R in a cross-lingual but not cross-domain transfer setup. For each domain, we first train an individual XLM-R model on domain-specific texts in English (see Appendix 4) for 100,000 update steps via Masked Language Modeling (MLM). After this intermediate domain fine-tuning, we fine-tune the domain-adapted models on the SQuAD 2.0 dataset.

LLaMA We evaluate the performance of Llama 2-chat 13B (Touvron et al., 2023)[8] and Llama 3-instruct 8B (AI@Meta, 2024). We apply simple postprocessing to extract the answer from the generated text; see Appendix D.3 for details.

GPT-3.5 We also experiment with GPT-3.5 (Brown et al., 2020). As its behavior changes over time (Chen et al., 2023), we investigate two versions of `gpt-3.5-turbo`: `-0301` and `-0613`.

Aya 23 Lastly, we evaluate Aya 23 8B (Aryabumi et al., 2024), a multilingual large language model.

4.2 Setup

Here we introduce a new modular setup that extends MAD-X for language and domain transfer. We propose two training variants: MAD-X+Domain and MAD-X^2. Appendix C provides details.

MAD-X+Domain We extend the MAD-X (Pfeiffer et al., 2020) language transfer framework with a domain adapter by stacking the task adapter above the domain adapter, which is stacked above the language adapter. We train new domain adapters and use MAD-X's language adapters that were trained via MLM on Wikipedia. Each domain adapter is trained for 100,000 update steps on the same English domain texts as XLM-R$^{\text{Base}}$ using MLM with an activated English language adapter. Then, we train the QA task adapter on SQuAD 2.0 with the

[8]Mostly trained on English (89.7% of the training data).

English language and Wikipedia domain adapter enabled. During evaluation, we activate the domain and language adapters of the target task.

MAD-X^2 The MAD-X^2 setup maintains the MAD-X+Domain's adapter architecture but alters the training approach to *simultaneously* train language and domain adapters. We use MLM on texts for every domain-language combination except for Chinese and Turkish creative writing where massive, open-licensed data is scarce. [9] During training, we change the domain and language to be trained in each batch, i.e., each batch has text from a different domain and language and the corresponding adapters are activated. We hypothesize that this fosters distinct encapsulation of language-specific and domain-specific knowledge within the respective adapters. We train every domain and language adapter for 62,500 update steps with a batch size of 16, resulting in a total of 250,000 update steps.

4.3 Results

We report the performance by language for the answerable and unanswerable questions of M2QA in Table 2, using the F1 and Exact Match (EM) scores as defined by Rajpurkar et al. (2018). For answerable questions, we report both scores, whereas for unanswerable questions, only the F1 score is included since both scores are identical by definition.

4.3.1 Performance of Existing Models (RQ1)

We first investigate how well existing models perform on the dataset. This includes LLMs and fine-tuned XLM-R baselines. We observe that out of all the approaches we evaluated, `gpt-3.5-turbo-0613` performs best with an average F1 score of 53.11 followed by Aya 23 with 51.61, `gpt-3.5-turbo-0301` with 47.08, Llama 3-instruct with 44.54.

Llama 2-chat (13b) with an average F1 score of 17.95 performs poorly in Turkish and Chinese, especially on the answerable questions. This is not surprising considering that LLama-2 is trained mainly on English text. In the zero-shot setting, Llama-2 often produces long responses that mix English with the target language. However, these issues are less pronounced in German, leading to comparatively better performance. To improve the performance of LLMs, we investigated using few-shot prompts. As detailed in Appendix D.1, only Llama 2 and `gpt-3.5-turbo-0301` consistently benefitted from this.

XLM-R$^{\text{Base}}$ has an average F1 score of 37.73 and performs well across languages, despite being smaller. XLM-R$^{\text{Domain}}$, with an average F1 score of 36.36, performs particularly poorly on answerable questions. This indicates that performing intermediate fine-tuning on English domain data is not only insufficient for domain transfer but actually hurts the performance, at least for German and Turkish. This is potentially caused by catastrophic forgetting of language-specific information (French, 1999).

To gain further insights into the performance of GPT-3.5, we manually inspected German questions for which all four GPT-3.5 setups achieved an F1 score lower than 25, which are 942 questions in total, or 20.9% of the German QA instances. We randomly sampled 50 questions from this subset to analyze the responses of the GPT-3.5 models. We found that in 72% of the cases, the question and answer are correctly annotated in the data, but the model either makes erroneous predictions (58%) or generates a correct answer instead of extracting it (14%). We further identified issues with inconsistent annotations (22%, i.e. 4.6% of all German data), questions with multiple plausible answers (4%), and the evaluation metric (2%). We detail this investigation in Appendix D.5.

4.3.2 Hard Domains and Languages (RQ2)

We now explore which languages and domains are particularly hard to tackle for the existing models. As per Table 2, for all explored models, the scores in German and Turkish are notably higher than the scores in Chinese, suggesting that this transfer is harder for the models. We revisit this observation in Section 5. Moreover, the performance in the news domain is higher than in creative writing and reviews. This shows that the model's domain transfer abilities still have room for improvement.

Performance on creative writing and product reviews varies by language. For German and Turkish, the results on product reviews are considerably better than on creative writing on the answerable questions, whereas in Chinese, the results are considerably better in creative writing for GPT-3.5 and Llama. This highlights the need for a joint evaluation of language and domain transfer. To investigate isolated cross-lingual and cross-domain transfer, we evaluated further setups, but could not

[9] The Chinese and Turkish creative writing test sets in M2QA were manually curated to ensure quality and exclude harmful content – yet sanitizing a large-scale corpus in such way was not feasible.

		Creative Writing			Product Reviews			News			Average				
		answerable		unansrbl	answerable		unansrbl	answerable		unansrbl	answerable		unansrbl	Total Average	
	Model	F1	EM	F1	F1	EM	F1	F1	EM	F1	F1	EM	F1	F1	EM
German	XLM-R^{Base}	30.42	17.67	67.83	35.64	21.22	56.67	40.98	26.56	55.33	35.68	21.82	59.94	45.38	37.07
	XLM-R^{Domain}	18.39	9.44	79.00	30.41	17.00	60.83	20.79	11.56	69.50	23.20	12.67	69.78	41.83	35.51
	MAD-X+Domain	4.25	2.44	**94.83**	23.44	13.56	73.33	38.82	24.44	55.33	22.17	13.48	74.50	43.19	37.98
	MAD-X^2	19.09	11.33	82.33	22.96	13.44	72.33	42.59	27.67	53.50	28.21	17.48	69.39	44.68	38.24
	Llama 2-chat (13b)	31.19	11.89	12.50	28.38	11.00	17.83	39.33	21.56	12.83	32.97	14.82	14.39	25.53	14.64
	Llama 3-instruct (8b)	**44.98**	**24.33**	34.33	45.19	**24.56**	32.17	55.41	37.44	30.33	**48.53**	**28.78**	34.65	42.98	31.13
	gpt-3.5-turbo-0301	40.49	20.67	64.67	**45.31**	24.00	61.17	**58.59**	36.22	58.17	48.13	26.96	61.34	53.41	40.71
	gpt-3.5-turbo-0613	37.68	22.22	80.50	42.22	24.44	76.50	55.53	37.67	**76.00**	45.14	28.11	77.67	**58.15**	**47.93**
	Aya-23 (8b)	31.69	19.89	82.17	35.82	20.89	**81.67**	55.30	**39.78**	74.00	40.94	26.85	**79.28**	56.28	47.82
Turkish	XLM-R^{Base}	22.65	14.78	68.50	32.68	17.44	59.33	41.71	29.56	57.17	32.35	20.59	61.67	44.08	37.02
	XLM-R^{Domain}	5.46	3.22	89.33	11.20	5.11	77.83	12.40	6.67	82.00	9.69	5.00	83.05	39.05	36.22
	MAD-X+Domain	2.15	1.33	96.00	11.33	6.00	90.33	30.97	20.78	66.17	14.82	9.37	84.17	42.64	39.04
	MAD-X^2	3.97	2.78	**97.17**	8.43	4.89	**93.17**	21.74	15.89	**83.50**	11.38	7.85	**91.28**	43.34	41.22
	Llama 2-chat (13b)	18.11	9.00	5.00	22.16	9.22	6.00	22.27	9.00	4.83	20.85	9.07	5.28	14.62	7.55
	Llama 3-instruct (8b)	**46.27**	**31.67**	25.17	54.06	28.56	36.50	53.91	32.67	26.50	51.41	30.97	29.39	**59.35**	30.34
	gpt-3.5-turbo-0301	36.58	20.33	68.33	53.63	25.00	60.83	53.67	27.44	54.50	47.96	24.26	61.22	53.26	39.04
	gpt-3.5-turbo-0613	44.26	28.56	75.17	**57.29**	**32.33**	63.67	**56.14**	**33.78**	57.67	**52.56**	**31.56**	65.50	57.74	**45.13**
	Aya-23 (8b)	41.74	31.11	65.83	52.93	30.78	59.83	52.59	32.56	54.33	49.09	31.48	60.00	53.45	42.89
Chinese	XLM-R^{Base}	0.11	0.11	32.33	0.69	0.56	35.67	39.67	24.44	49.33	13.49	8.37	39.11	23.74	20.67
	XLM-R^{Domain}	0.00	0.00	48.17	0.28	0.22	62.33	1.79	1.00	**98.00**	0.69	0.41	69.50	28.21	28.05
	MAD-X+Domain	0.00	0.00	**92.00**	0.39	0.33	**85.00**	32.21	20.22	60.17	10.87	6.85	**79.06**	38.43	36.02
	MAD-X^2	0.11	0.11	79.67	0.17	0.11	83.67	33.24	22.00	68.67	11.17	7.41	77.34	37.64	35.38
	Llama 2-chat (13b)	13.05	2.44	16.17	12.39	3.89	17.50	10.86	1.89	14.67	12.10	2.74	16.11	13.70	8.09
	Llama 3-instruct (8b)	**36.61**	33.33	29.17	**27.52**	23.33	32.17	33.50	21.00	26.83	32.54	25.89	29.39	31.28	27.29
	gpt-3.5-turbo-0301	27.12	24.78	57.00	19.50	16.56	60.50	18.86	15.00	43.50	21.83	18.78	53.67	34.56	32.73
	gpt-3.5-turbo-0613	35.31	34.44	66.50	26.01	25.44	71.67	27.88	21.78	53.83	29.73	27.22	64.00	43.44	**41.93**
	Aya-23 (8b)	35.44	**35.44**	56.83	26.74	**26.74**	65.33	**50.39**	**33.67**	47.33	**37.52**	**31.95**	56.50	**45.11**	41.77

Table 2: Results of the base models and adapter-based methods on the M2QA benchmark using the F1/EM score definitions by SQuAD 2.0 (Rajpurkar et al., 2018). See our discussion of this metric's potential flaws in Section 5.1. For the answerable questions, we report the F1 and Exact Match (EM) scores. For the unanswerable (unansrbl) questions, we only include the F1 score as the EM score is identical to it by definition. The average is taken across datapoints. The best score for each language in each column is bold.

find improved performance (Appendix D.4).

4.3.3 Modular setups (RQ3)

Finally, we use M2QA to evaluate our two modular adaptation setups: MAD-X+Domain and MAD-X^2. Based on our results (Table 2), these setups achieve average scores on par with XLM-R^{Base} in German and Turkish, and notably improve the Chinese score. We note that this increase primarily stems from the improved performance on unanswerable questions, while the performance on answerable questions declines. Despite similar overall performance between MAD-X+Domain and MAD-X^2, a notable difference lies in the number of update steps during training: MAD-X+Domain was trained a total of 1M training steps (100k for each domain and 250k for each language, with a batch size of 64), while MAD-X^2 only needs 250k training steps with batch size 16 to achieve similar performance. This highlights MAD-X^2 computational efficiency, indicating the potential for simultaneous training of language and domain adapters.

5 Further Analysis

In contrast to English, German, and Turkish, which use whitespace characters to separate words, in Chinese typesetting the use of whitespace is not *required*. While the texts from our Chinese product review and creative writing sources do not contain whitespaces, Chinese news do. We hypothesize that this typographical difference between Chinese and the other languages can lead to a substantial drop in measured performance (e.g. XLM-R^{Base} achieves an F1 score of 0.11 on answerable creative writing questions), and investigate this further.

5.1 SQuAD Metric - Adaptation for Chinese

For the evaluation in Section 4.3, we have used the F1/EM definitions of SQuAD 2.0, which is widely adopted and has been previously used to evaluate multilingual extractive QA (e.g. Artetxe et al., 2020b). During the metric calculation, this implementation splits words by whitespace – however, if whitespaces are not available, the whole text is

considered as one long token, rendering the rest of the calculation invalid. We modify the implementation to make the metric applicable to Chinese texts without whitespace tokenization by splitting the text into tokens using the off-the-shelf jieba tokenizer[10]. The resulting measurements, detailed in Appendix E.1, differ substantially from those in Table 2, suggesting that the SQuAD metric implementation needs adjustment for multilingual extractive QA evaluation. Even for texts from the news domain which contain whitespace, the tokenizer-based version of the metric results in higher scores. The tokenizer splits the Chinese text into smaller tokens than whitespace tokenization, allowing a finer-grained score. Moreover, the XLM-R-based methods struggle to make meaningful predictions for text without whitespace (see Section 5.2). Since the score only improves for spans close to the gold span, the improvement for LLMs is bigger than for XLM-R-based methods.

5.2 Adding Whitespaces to Chinese Text

Having examined the predictions of the XLM-R-based methods, we found that training on English SQuAD data leads to XLM-R returning spans surrounded by whitespace as answers. If the Chinese text does not contain whitespaces, XLM-R-based methods either classify the question as unanswerable or return the whole passage as the answer. To explore the impact of this issue, we re-run the XLM-R$^{\text{Base}}$ setup but added whitespace to the texts between jieba-determined words.

This modification leads to improved performance on Chinese texts with no whitespace (+24.9 F1 points for creative writing, +17.7 F1 points for product reviews) but reduces the measured performance on texts with whitespace (-7.1 F1 points for news). The improved performance on texts that previously had no whitespace suggests that language transfer methods like MAD-X struggle to transfer tasks to languages without inherent whitespace. The reduced performance on texts that already contained whitespace indicates that adding whitespace between jieba-determined words is not yet optimal. This suggests that typographical features of the source data can affect measured performance and should be taken into account when experimenting with non-Latin-based languages. Heuristics, i.e. whitespaces added through tokenization, can help improve performance. Detailed results are in Appendix E.2

[10] https://github.com/fxsjy/jieba v0.42.1

6 Discussion and Future Work

M2QA allows us to evaluate joint language and domain transfer across different language models and adaptation approaches. Our results indicate room for improvement, especially when comparing the results of XLM-R-based models and LLMs. Since 40% of M2QA's questions are unanswerable, a naive model that classifies all questions as unanswerable would reach an F1/EM score of 40.0/40.0. For Chinese, only `gpt-3.5-turbo-0613` and Aya 23 perform better than this naive strategy, emphasizing the need for more sophisticated domain and language transfer methods. We hope that our resource enables and encourages work on systematically exploring prompts and setups that perform transfer learning across multiple dimensions of language variation. Future efforts should also aim to add more languages and domains to M2QA, especially for low-resource languages and domains. We hope that our published annotation protocols and software will facilitate this work[11]. Finally, establishing human performance baselines would help us understand how far NLP systems are from achieving human-level extractive QA performance across languages and domains.

7 Conclusion

Generalization is a central goal of NLP that is yet unsolved. Language and domain are two main axes of variation for natural languages – yet the lack of cross-lingual cross-domain datasets has prevented systematic evaluation of NLP models and transfer approaches across languages and domains. To address this, we introduce M2QA, a multi-domain multilingual question answering benchmark with over 13k human-annotated instances across three typologically diverse languages (German, Turkish, Chinese) and three distinct domains (product reviews, news, creative writing). Our evaluation includes XLM-R baselines, LLMs (GPT-3.5, Aya 23, Llama 2 and 3), and adapter-based setups (MAD-X+Domain and MAD-X^2), revealing a large gap between LLMs and fine-tuned LMs. We expect that M2QA will help close this gap, increase our understanding of generalization, and find more effective domain and language transfer methods.

[11] The code for all experiments, including hyperparameters, prompts, and the implementation of the annotation environment, is available in our GitHub repository.

8 Limitations

A major obstacle to including more languages and domains into M2QA has been a *severe shortage* of *clearly and openly licensed* unlabeled texts in under-represented language-domain combinations – due to the restrictive copyright in many domains (news, books), and due to the lack of explicit licensing practices in others. While we made an effort to diversify the selection of languages and domains in M2QA, the dataset only covers a small subset of all existing languages and domains. Potential solutions to this could be to clarify or obtain a license for research use from the owners of the textual data, as well as to experiment with data synthetically generated via paraphrasing or machine translation. However, translations introduce considerable issues, including lowered lexical diversity, "translationese", and lack of cultural idiosyncrasies, as discussed in the introduction and Section 3.1. This exploration, as well as the comparison between the results on synthetic and natural QA data, is left to the future.

Since some of the data sources in M2QA are widely used (e.g. Gutenberg Corpus or Amazon Reviews), there is a risk that LLMs have observed some of the unlabeled data during their pre-training. The unavailability of pre-training data for LLaMa 2, Llama 3, GPT-3.5 and Aya 23 prevents us from investigating whether this is the case. To prevent contamination of future experimental setups with the labelled data, we employ protective measures, following Jacovi et al. (2023): We release the data in encrypted form with a CC-BY-ND 4.0 license.[12]

We evaluate with XLM-R for a consistent setup with MAD-X (Pfeiffer et al., 2020).

References

AI@Meta. 2024. Llama 3 model card.

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. 2022. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1796, Dublin, Ireland. Association for Computational Linguistics.

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020a. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

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

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. 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 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Elisa Bassignana, Filip Ginter, Sampo Pyysalo, Rob van der Goot, and Barbara Plank. 2023. MultiCrossRE a multi-lingual multi-domain dataset for relation extraction. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 80–85, Tórshavn, Faroe Islands. University of Tartu Library.

Elisa Bassignana and Barbara Plank. 2022. CrossRE: A cross-domain dataset for relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3592–3604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, Dublin, Ireland. Association for Computational Linguistics.

Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese?

[12] https://creativecommons.org/licenses/by-nd/4.0/legalcode

comparing human and machine translations of text and speech. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 280–290, Online. Association for Computational Linguistics.

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A Smith, and Luke Zettlemoyer. 2024. Breaking the curse of multilinguality with cross-lingual expert language models. *arXiv*, arXiv preprint arXiv:2401.10440.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is chatgpt's behavior changing over time? *arXiv preprint arXiv:2307.09009*.

Alexandra Chronopoulou, Matthew Peters, and Jesse Dodge. 2022. Efficient hierarchical domain adaptation for pretrained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1336–1351, Seattle, United States. Association for Computational Linguistics.

Alexandra Chronopoulou, Matthew Peters, Alexander Fraser, and Jesse Dodge. 2023. AdapterSoup: Weight averaging to improve generalization of pretrained language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2054–2063, Dubrovnik, Croatia. 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. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

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

Asa Cooper Stickland, Alexandre Berard, and Vassilina Nikoulina. 2021. Multilingual domain adaptation for NMT: Decoupling language and domain information with adapters. In *Proceedings of the Sixth Conference on Machine Translation*, pages 578–598, Online. Association for Computational Linguistics.

Benjamin D. Douglas, Patrick J. Ewell, and Markus Brauer. 2023. Data quality in online human-subjects research: Comparisons between MTurk, prolific, CloudResearch, qualtrics, and SONA. *Plos one*, 18(3):e0279720. Publisher: Public Library of Science.

Shiran Dudy and Steven Bedrick. 2020. Are some words worth more than others? In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 131–142, Online. Association for Computational Linguistics.

Frank Fischer and Jannik Strötgen. 2015. When does (german) literature take place? on the analysis of temporal expressions in large corpora. In *Proceedings of DH 2015: Annual Conference of the Alliance of Digital Humanities Organizations*, volume 6.

Wikimedia Foundation. Wikimedia downloads.

Robert M. French. 1999. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3:128–135.

Martin Gerlach and Francesc Font-Clos. 2018. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, 22.

Abbas Ghaddar and Phillippe Langlais. 2017. WiNER: A Wikipedia annotated corpus for named entity recognition. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 413–422, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Mandy Guo, Yinfei Yang, Daniel Cer, Qinlan Shen, and Noah Constant. 2021. MultiReQA: A cross-domain evaluation forRetrieval question answering models. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 94–104, Kyiv, Ukraine. Association for Computational Linguistics.

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 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Shivanshu Gupta, Yoshitomo Matsubara, Ankit Chadha, and Alessandro Moschitti. 2023. Cross-lingual knowledge distillation for answer sentence selection in low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14078–14092, Toronto, Canada. Association for Computational Linguistics.

Suchin Gururangan, Mike Lewis, Ari Holtzman, Noah A. Smith, and Luke Zettlemoyer. 2022. DEMix layers: Disentangling domains for modular language modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5557–5576, Seattle, United States. Association for Computational Linguistics.

Suchin Gururangan, Margaret Li, Mike Lewis, Weijia Shi, Tim Althoff, Noah A. Smith, and Luke Zettlemoyer. 2023. Scaling expert language models with unsupervised domain discovery. *arXiv*, arXiv preprint arXiv: 2303.14177.

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*, pages 8342–8360, Online. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

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, Melbourne, Australia. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

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. *arXiv*, arXiv preprint arXiv: 2003.11080.

Songbo Hu, Han Zhou, Mete Hergul, Milan Gritta, Guchun Zhang, Ignacio Iacobacci, Ivan Vulić, and Anna Korhonen. 2023a. Multi 3 WOZ: A multilingual, multi-domain, multi-parallel dataset for training and evaluating culturally adapted task-oriented dialog systems. *Transactions of the Association for Computational Linguistics*, 11:1396–1415.

Songbo Hu, Han Zhou, Mete Hergul, Milan Gritta, Guchun Zhang, Ignacio Iacobacci, Ivan Vulic, and Anna Korhonen. 2023b. Multi3woz: A multilingual, multi-domain, multi-parallel dataset for training and evaluating culturally adapted task-oriented dialog systems. *CoRR*, abs/2307.14031.

Gabriel Ilharco, Marco Túlio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084, Singapore. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

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

Mayank Kulkarni, Daniel Preotiuc-Pietro, Karthik Radhakrishnan, Genta Indra Winata, Shijie Wu, Lingjue Xie, and Shaohua Yang. 2023. Towards a unified multi-domain multilingual named entity recognition model. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2210–2219, Dubrovnik, Croatia. Association for Computational Linguistics.

Hele-Andra Kuulmets and Mark Fishel. 2023. Translated benchmarks can be misleading: the case of Estonian question answering. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 710–716, Tórshavn, Faroe Islands. University of Tartu Library.

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

Wen Lai, Alexandra Chronopoulou, and Alexander Fraser. 2022. m^4 adapter: Multilingual multi-domain adaptation for machine translation with a meta-adapter. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4282–4296, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

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*, pages 7315–7330, Online. Association for Computational Linguistics.

Margaret Li, Suchin Gururangan, Tim Dettmers, Mike Lewis, Tim Althoff, Noah A. Smith, and Luke Zettlemoyer. 2022. Branch-train-merge: Embarrassingly parallel training of expert language models. *arXiv preprint*, abs/2208.03306.

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. *arXiv preprint*, abs/1907.11692.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2023. Are emergent abilities in large language models just in-context learning? *arXiv preprint*, abs/2309.01809.

Benjamin Minixhofer, Jonas Pfeiffer, and Ivan Vulić. 2023. Where's the point? self-supervised multilingual punctuation-agnostic sentence segmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7215–7235, Toronto, Canada. Association for Computational Linguistics.

Nikita Moghe, Evgeniia Razumovskaia, Liane Guillou, Ivan Vulić, Anna Korhonen, and Alexandra Birch. 2023. Multi3NLU++: A multilingual, multi-intent, multi-domain dataset for natural language understanding in task-oriented dialogue. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3732–3755, Toronto, Canada. Association for Computational Linguistics.

Tarek Naous, Michael J. Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2023. Readme++: Benchmarking multilingual language models for multi-domain readability assessment. *arXiv preprint*, abs/2305.14463.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint*, abs/2303.08774.

Marinela Parovic, Alan Ansell, Ivan Vulić, and Anna Korhonen. 2023. Cross-lingual transfer with target language-ready task adapters. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 176–193, Toronto, Canada. Association for Computational Linguistics.

Marinela Parović, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2022. BAD-X: Bilingual adapters improve zero-shot cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1791–1799, Seattle, United States. Association for Computational Linguistics.

Jonas Pfeiffer, Sebastian Ruder, Ivan Vulić, and Edoardo Ponti. 2023. Modular deep learning. *Transactions on Machine Learning Research*. Survey Certification.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Clifton Poth, Jonas Pfeiffer, Andreas Rücklé, and Iryna Gurevych. 2021. What to pre-train on? Efficient intermediate task selection. In *Proceedings of the*

2021 Conference on Empirical Methods in Natural Language Processing, pages 10585–10605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Clifton Poth, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vulić, Sebastian Ruder, Iryna Gurevych, and Jonas Pfeiffer. 2023. Adapters: A unified library for parameter-efficient and modular transfer learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 149–160, Singapore. Association for Computational Linguistics.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.

Ella Rabinovich, Sergiu Nisioi, Noam Ordan, and Shuly Wintner. 2016. On the similarities between native, non-native and translated texts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1881, Berlin, Germany. Association for Computational Linguistics.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. 2020. Compressive transformers for long-range sequence modelling. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

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*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Machel Reid, Victor Zhong, Suchin Gururangan, and Luke Zettlemoyer. 2022. M2D2: A massively multi-domain language modeling dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 964–975, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting closer to ai complete question answering: A set of prerequisite real tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8722–8731.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2022. Square one bias in NLP: Towards a multi-dimensional exploration of the research manifold. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2340–2354, Dublin, Ireland. Association for Computational Linguistics.

Dietmar Schabus, Marcin Skowron, and Martin Trapp. 2017. One million posts: A data set of german online discussions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 1241–1244. ACM.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 2214–2218. European Language Resources Association (ELRA).

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020. Label studio: Data labeling software.

Cagri Toraman, Fazli Can, and Seyit Koçberber. 2011. Developing a text categorization template for turkish news portals. In *2011 International Symposium on Innovations in Intelligent Systems and Applications*, pages 379–383.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint*, abs/2307.09288.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *arXiv preprint*, abs/2306.07899.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Danqing Wang, Jiaze Chen, Xianze Wu, Hao Zhou, and Lei Li. 2021. Cnewsum: A large-scale summarization dataset with human-annotated adequacy and deducibility level. In *Natural Language Processing and Chinese Computing - 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13-17, 2021, Proceedings, Part I*, volume 13028 of *Lecture Notes in Computer Science*, pages 389–400. Springer.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 2023. Should you mask 15% in masked language modeling? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2985–3000, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Steve Yadlowsky, Lyric Doshi, and Nilesh Tripuraneni. 2023. Pretraining data mixtures enable narrow model selection capabilities in transformer models. *arXiv preprint*, abs/2311.00871.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 19–27. IEEE Computer Society.

A Passage Curation

Table 3 shows the datasets we selected and their licensing information. The final dataset should contain 300 passages for each language and domain combination. To preprocess the data and prepare the passages, we need to distinguish between the domains that have multiple passages per document and the ones with one passage per document. For ones where a document is one passage, we filter out documents that are too short and too long. From the remaining documents, we randomly sample 300. For the domains that have multiple passages per document, we first exclude the ones that are too short to feature at least three passages. Then, we sample documents and split them into passages using the WTP segmentation model (Minixhofer et al., 2023). We use only documents with at least three passages. The German creative writing of the Gutenberg corpus required a different setup. Because of the different formatting of footnotes, references, and diverse formatting of bold, underlined, and cursive text, we manually extracted 300 passages from 6 fiction creative writing that had licenses that made them free to use. The passages from all domains are then stripped of newline characters, tabs, and multiple consecutive white spaces. The creative writing passages for Turkish and Chinese are taken from an online social reading platform where people can publish their own stories. We select texts published in the public domain or with a Creative Common License. To ensure that no author's notes or unsuitable or offensive texts, such as comments or sensitive topics, are in the passage, we manually check the translated [13] passages.

A.1 Lexical Diversity of the Domains

To quantify lexical diversity in the data, in Figure 2, we report the Jaccard similarity coefficient of the vocabularies between the different domains in one language.[14] As we observe, the domains in M2QA indeed show low vocabulary overlaps, making our dataset a challenging target for domain adaptation across languages.

B Baseline Training

All our models are based on XLM-R-base (Conneau et al., 2020), a multilingual 270M parameter model. XLM-R$^{\text{Base}}$ is directly fine-tuned on SQuAD 2.0, while XLM-R$^{\text{Domain}}$ has been domain-adapted prior to fine-tuning on SQuAD 2.0 (Rajpurkar et al., 2018). Every not-mentioned hyper-

[13] We use DeepL for translation.

[14] We use nltk `https://www.nltk.org` for German and Turkish tokenization, and jieba `https://github.com/fxsjy/jieba` for Chinese

Language	Domain	Multiple Passages	Datasource	License
German	product reviews	no	Amazon Reviews (Keung et al., 2020)	Usage permitted by Amazon for academic research[1].
	news	yes	10kGNAD[2]	CC BY-NC-SA 4.0
	creative writing	yes	Gutenberg Corpus (Gerlach and Font-Clos, 2018)	Manually selected text passages from open-license books.
Turkish	product reviews	no	Turkish product reviews[3]	CC BY-SA 4.0
	news	yes	BilCat (Toraman et al., 2011)	MIT License
	creative writing	yes	Wattpad[4]	Manually selected text passages from Creative Commons or Public Domain publications.
Chinese	product reviews	no	Amazon Reviews (Keung et al., 2020)	Usage permitted by Amazon for academic research[1].
	news	yes	CNewSum (Wang et al., 2021)	MIT License
	creative writing	yes	Wattpad[4]	Manually selected text passages from Creative Commons or Public Domain publications.

Table 3: The original datasets used for annotation

[1] https://github.com/awslabs/open-data-docs/blob/main/docs/amazon-reviews-ml/license.txt
[2] https://github.com/tblock/10kGNAD using the One Million Posts dataset by Schabus et al. (2017)
[3] https://huggingface.co/datasets/turkish_product_reviews
[4] https://www.wattpad.com/

Domain	Datasource
Wikipedia	Wikipedia (Foundation)
Creative Writing	bookcorpus (Zhu et al., 2015)
Product Reviews	Amazon Reviews (Keung et al., 2020)
News	CNN Dailymail (Hermann et al., 2015)

Table 4: Texts used for adapting the domain of XLM-R$^{\text{Domain}}$ and the domain adapters of MAD-X+Domain.

parameter is the default parameter of Hugging Face Transformers (Wolf et al., 2020) version 4.26.1.

XLM-R$^{\text{Base}}$ We train XLM-R on SQuAD 2.0 for 100,000 update steps, use early stopping with patience of 5 and evaluate every 1000 steps. We use a batch size of 64, 1000 warmup steps, a learning rate of 1e-4, linear learning rate decay and a sequence length of 512.

XLM-R$^{\text{Domain}}$ We first train an individual model for each domain via MLM on the data displayed in Table 4. We train for 100,000 update steps with a batch size of 16, a learning rate of 1e-4, linear learning rate decay and a sequence length of 512. Following Wettig et al. (2023), we use an MLM probability of 40% since XLM-R-base has a comparable size to bert-large. After this training, we fine-tune every domain-adapted XLM-R model on SQuAD 2.0 with the same parameters used for XLM-R$^{\text{Base}}$.

C MAD-X+Domain & MAD-X^2 Training

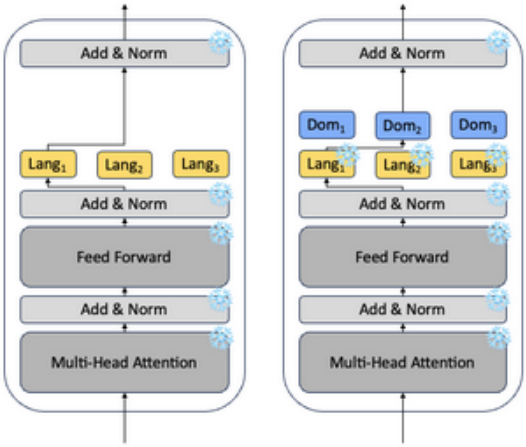
We use the *Adapters* library (Poth et al., 2023) for the adapter and model implementations.

MAD-X+Domain The domain adapters are Pfeiffer Bottleneck Adapters with a reduction factor of 2. We train each domain adapter for 100,000 update steps with a batch size of 16, 1000 warmup steps, learning rate of 1e-4 and linear learning rate decay via masked language modelling on English data. The data sources used for each adapter are listed in Table 4. Since we train on English data, we activate the English MAD-X (Pfeiffer et al., 2020) adapter. Following Wettig et al. (2023), we use an MLM probability of 40% since XLM-R-base has a comparable size to bert-large. Overall, the domain and language adapters cumulatively used 1,400,000 update steps with a batch size of 64 (the four language adapters were trained with 250,000 steps, the four domain adapters with 100,000).

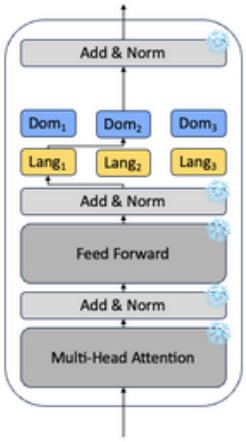
The QA head adapter, also a Pfeiffer Bottleneck Adapter with a reduction factor of 2, was trained on SQuADv2 (Rajpurkar et al., 2018) using the same hyper-parameters used for fine-tuning XLM-R$^{\text{Base}}$. Since SQuADv2 is based on English Wikipedia text passages, we activated the English and Wikipedia adapter during training.

(a) German (b) Turkish (c) Chinese

Figure 2: Jaccard similarity coefficient of the vocabularies between the domains, per language.



(a) MAD-X+Domain: The language adapter is trained as proposed in the MAD-X setup, and the domain adapter is trained in a second step with the corresponding frozen language adapter activated.



(b) MAD-X^2: During training, the language and domain adapters are trained simultaneously, and each training sample is routed through the corresponding language and domain adapter.

Figure 3: The different pre-training setups for MAD-X+Domain, and MAD-X^2

MAD-X^2 We use the same hyper-parameters as for the training of the domain adapters of MAD-X+Domain. The only parameter changed is the number of update steps where we train every domain and language adapter for 62,500 update steps, resulting in a total of 250,000 update steps. The corpora used for the MLM training are listed in Table 5 along with the number of steps trained on each corpus. Due to the absence of open-license text corpora, we do not train on Chinese and Turkish creative writing corpora. The QA head adapter is trained afterwards identical to the QA head adapter of MAD-X+Domain.

D LLM Evaluation

D.1 Five-Shot LLM Results

We explored if providing a few-shot prompt could enhance the performance of the LLMs. In Table 2, we present the results of the LLMs using a five-shot prompt. Upon comparing these results with the zero-shot results in Table 6, we see that the five-shot prompt does not consistently improve performance. The only models benefitting from the five-shot prompt are Llama 2 and `gpt-3.5-turbo-0301`.

D.2 LLM Prompts

Based on the extractive question answering prompt of Lai et al. (2023), we evaluate `gpt-3.5-turbo-0301`, `gpt-3.5-turbo-0613` and LLaMA 2-chat 13B in a zero-shot and five-shot setting. The zero-shot prompt is displayed in Figure 4. The five-shot prompts contain three answerable and two unanswerable examples. Following Brown et al. (2020), we provide the five examples in a single user prompt, as shown in Figure 5. However, this setup yielded scores close to zero for Llama 2-chat 13B. Hence, we changed the Llama 2-chat

Language	Domain	Datasource	steps trained
English	Wikipedia	Wikipedia (Foundation)	10417
	Creative Writing	PG-19 (Rae et al., 2020)	31250
	Product Reviews	Amazon Reviews (Keung et al., 2020)	10416
	News	CNN Dailymail (Hermann et al., 2015)	10417
German	Wikipedia	Wikipedia (Foundation)	10417
	Creative Writing	Opus Books (Tiedemann, 2012) & Corpus of German-Language Fiction (Fischer and Strötgen, 2015)	31250
	Product Reviews	Amazon Reviews (Keung et al., 2020)	10416
	News	MLSUM (Scialom et al., 2020)	10417
Turkish	Wikipedia	Wikipedia (Foundation)	20833
	Creative Writing		0
	Product Reviews	Turkish Product Reviews[1]	20833
	News	XL-Sum (Hasan et al., 2021)	20833
Chinese	Wikipedia	Wikipedia (Foundation)	20833
	Creative Writing		0
	Product Reviews	Amazon Reviews (Keung et al., 2020)	20833
	News	XL-Sum (Hasan et al., 2021)	20833

Table 5: MAD-X² training data sources & number of steps trained

[1] https://huggingface.co/datasets/turkish_product_reviews

System Prompt:
Task Description: Answer the question from the given passage. Your answer should be directly extracted from the passage, and it should be a single entity, name, or number, not a sentence. If the passage doesn't contain a suitable answer, please respond with 'unanswerable'.

User:
Passage: {CONTEXT}
Question: {QUESTION}
Note: Your answer should be directly extracted from the passage and be a single entity, name, or number, not a sentence. If the passage doesn't contain a suitable answer, respond with 'unanswerable'.
Answer:

Figure 4: Zero-shot English Prompt

13B setup by providing examples not in a single prompt but as part of the chat history.

D.3 LLM Postprocessing

In the zero-shot setting, Llama tends to generate more than just the answer span; for instance, "Based on the passage, the answer to the question is: [...] The passage states: [...]". This is not the only pattern in which the model phrases the answer. To capture as many as possible, we split the text at the semicolon and take the part that follows the semicolon. The answer and potential text passages to back it up are, in most cases, separated by newlines. We use this to remove text that is not part of the answer span: If there is no semicolon, we take the whole text output as the answer. These problems are particularly pronounced with Llama 2 and occur less with Llama 3.

D.4 Isolated Domain and Language Transfer

To further investigate the effect of domain and language, we investigate domain transfer and language transfer isolated. Overall, these configurations did not improve performance.

Isolated Language Transfer To explore the isolated language transfer, we eliminated domain variation. We provide GPT-3.5 with a prompt in a different language and examples in the language from the same domain. German gets a Turkish prompt, Turkish gets a Chinese prompt and Chinese gets a German prompt. This results in no improved performance as can be seen in Table 7.

Isolated Domain Transfer To perform only domain transfer, we eliminate language variation in the chat. We let native speakers translate the prompts to the target languages (German, Turkish, Chinese). Thus, in the zero-shot scenario, the model gets the system prompt, passage, question and note in the target language. For the five-shot evaluation, the examples come from the same language but from a different domain. The results are shown in Table 8.

D.5 Investigating German GPT-3.5 Answers

To gain further insights into GPT-3.5's performance, we chose to sample some hard questions and include a case study to analyze them. We manually inspected German questions for which all four GPT-3.5 setups achieved an F1 score lower than 25, which are 942 questions in total (20.9% of all German QA instances). From

System Prompt:
Task Description: Answer the question from the given passage. Your answer should be directly extracted from the passage, and it should be a single entity, name, or number, not a sentence. If the passage doesn't contain a suitable answer, please respond with 'unanswerable'.

User:
Passage: In 2007, BSkyB and Virgin Media became involved in a dispute over the carriage of Sky channels on cable TV. The failure to renew the existing carriage agreements negotiated with NTL and Telewest resulted in Virgin Media removing the basic channels from the network on 1 March 2007. Virgin Media claimed that BSkyB had substantially increased the asking price for the channels, a claim which BSkyB denied, on the basis that their new deal offered "substantially more value" by including HD channels and Video On Demand content which was not previously carried by cable.
Question: What channels were removed from the network in March of 2007?
Note: Your answer should be directly extracted from the passage and be a single entity, name, or number, not a sentence. If the passage doesn't contain a suitable answer, respond with 'unanswerable'.
Answer: the basic channels

Passage: Following the Cretaceous–Paleogene extinction event, the extinction of the dinosaurs and the wetter climate may have allowed the tropical rainforest to spread out across the continent. From 66–34 Mya, the rainforest extended as far south as 45°. Climate fluctuations during the last 34 million years have allowed savanna regions to expand into the tropics. During the Oligocene, for example, the rainforest spanned a relatively narrow band. It expanded again during the Middle Miocene, then retracted to a mostly inland formation at the last glacial maximum. However, the rainforest still managed to thrive during these glacial periods, allowing for the survival and evolution of a broad diversity of species.
Question: Savannah areas expanded over the last how many years?
Note: Your answer should be directly extracted from the passage and be a single entity, name, or number, not a sentence. If the passage doesn't contain a suitable answer, respond with 'unanswerable'.
Answer: 34 million

Passage: It is conjectured that a progressive decline in hormone levels with age is partially responsible for weakened immune responses in aging individuals. Conversely, some hormones are regulated by the immune system, notably thyroid hormone activity. The age-related decline in immune function is also related to decreasing vitamin D levels in the elderly. As people age, two things happen that negatively affect their vitamin D levels. First, they stay indoors more due to decreased activity levels. This means that they get less sun and therefore produce less cholecalciferol via UVB radiation. Second, as a person ages the skin becomes less adept at producing vitamin D.
Question: As a person gets older, what does the skin produce less of?
Note: Your answer should be directly extracted from the passage and be a single entity, name, or number, not a sentence. If the passage doesn't contain a suitable answer, respond with 'unanswerable'.
Answer: vitamin D

Passage: In 1066, Duke William II of Normandy conquered England killing King Harold II at the Battle of Hastings. The invading Normans and their descendants replaced the Anglo-Saxons as the ruling class of England. The nobility of England were part of a single Normans culture and many had lands on both sides of the channel. Early Norman kings of England, as Dukes of Normandy, owed homage to the King of France for their land on the continent. They considered England to be their most important holding (it brought with it the title of King—an important status symbol).
Question: What battle took place in the 10th century?
Note: Your answer should be directly extracted from the passage and be a single entity, name, or number, not a sentence. If the passage doesn't contain a suitable answer, respond with 'unanswerable'.
Answer: unanswerable

Passage: Dendritic cells (DC) are phagocytes in tissues that are in contact with the external environment; therefore, they are located mainly in the skin, nose, lungs, stomach, and intestines. They are named for their resemblance to neuronal dendrites, as both have many spine-like projections, but dendritic cells are in no way connected to the nervous system. Dendritic cells serve as a link between the bodily tissues and the innate and adaptive immune systems, as they present antigens to T cells, one of the key cell types of the adaptive immune system.
Question: What is named for its resemblance to dendritic cells?
Note: Your answer should be directly extracted from the passage and be a single entity, name, or number, not a sentence. If the passage doesn't contain a suitable answer, respond with 'unanswerable'.
Answer: unanswerable

Passage: {CONTEXT}
Question: {QUESTION}
Note: Your answer should be directly extracted from the passage and be a single entity, name, or number, not a sentence. If the passage doesn't contain a suitable answer, respond with 'unanswerable'.
Answer:

Figure 5: Five-shot English Prompt using SQuAD 2.0 examples.

	Model		Creative Writing answerable F1	Creative Writing answerable EM	Creative Writing unansrbl F1	Product Reviews answerable F1	Product Reviews answerable EM	Product Reviews unansrbl F1	News answerable F1	News answerable EM	News unansrbl F1	Average answerable F1	Average answerable EM	Average unansrbl F1	Average Total Average F1	Average Total Average EM
German	Llama 2-chat (13b)	(5-shot)	22.61	12.33	75.17	20.52	12.67	75.33	29.33	19.89	77.33	24.15	14.96	75.94	44.87 (+19.34)	39.36 (+24.72)
	Llama 3-instruct (8b)	(5-shot)	**43.44**	**23.89**	35.50	39.82	20.56	25.50	54.71	35.56	28.00	45.99	26.67	29.67	39.46 (-3.52)	27.87 (-3.26)
	gpt-3.5-turbo-0301	(5-shot)	40.00	21.89	76.83	**47.81**	**24.33**	60.67	**61.18**	38.22	59.67	**49.66**	28.15	65.72	56.09 (+2.68)	43.18 (+2.47)
	gpt-3.5-turbo-0613	(5-shot)	34.97	22.67	83.33	40.36	23.33	79.17	58.50	**39.44**	70.83	44.61	**28.48**	77.78	**57.88** (-0.27)	**48.20** (+0.27)
	Aya-23 (8b)	(5-shot)	16.44	10.78	**90.83**	19.85	12.11	**93.67**	38.00	28.44	**89.33**	24.76	17.11	**91.28**	51.37 (-4.91)	46.78 (-1.04)
Turkish	Llama 2-chat (13b)	(5-shot)	7.18	5.22	**84.83**	8.48	4.78	**91.00**	9.78	6.33	**88.17**	8.48	5.44	**88.00**	40.29 (+25.67)	38.47 (+30.92)
	Llama 3-instruct (8b)	(5-shot)	**38.73**	**24.33**	47.33	41.27	21.89	42.33	51.15	30.22	25.83	43.72	25.48	38.50	41.63 (-17.72)	30.69 (+0.35)
	gpt-3.5-turbo-0301	(5-shot)	36.82	22.78	73.83	**53.89**	24.33	61.67	**58.13**	**32.33**	55.00	**49.61**	**26.48**	63.50	55.17 (+1.91)	41.29 (+2.25)
	gpt-3.5-turbo-0613	(5-shot)	33.58	20.89	84.50	49.98	**25.78**	71.17	55.21	32.00	58.00	46.26	26.22	71.22	**56.24** (-1.5)	44.22 (-0.91)
	Aya-23 (8b)	(5-shot)	28.52	21.89	79.33	33.14	20.56	84.67	32.95	20.44	74.00	31.54	20.96	79.33	50.66 (-2.79)	**44.31** (+1.42)
Chinese	Llama 2-chat (13b)	(5-shot)	0.71	0.67	**95.33**	1.50	1.44	**90.83**	1.19	0.78	**96.00**	1.13	0.96	**94.05**	38.30 (+24.6)	38.20 (+30.11)
	Llama 3-instruct (8b)	(5-shot)	**34.78**	**32.22**	34.83	21.16	17.67	32.67	33.21	20.67	22.83	29.72	23.52	30.11	29.88 (-1.4)	26.16 (–1.13)
	gpt-3.5-turbo-0301	(5-shot)	29.19	26.22	63.17	21.05	16.89	63.67	21.55	17.33	48.50	23.93	20.15	58.45	37.74 (+3.18)	35.47 (+2.74)
	gpt-3.5-turbo-0613	(5-shot)	29.30	28.33	75.83	16.07	14.67	82.33	25.44	20.11	57.83	23.60	21.04	72.00	42.96 (-0.48)	**41.42** (-0.51)
	Aya-23 (8b)	(5-shot)	28.11	28.11	60.33	**22.74**	**22.67**	61.17	**50.40**	**34.33**	50.83	**33.75**	**28.37**	57.44	**43.23** (-1.88)	40.00 (-1.77)

Table 6: Results of the LLMs with five-shot prompts on the M2QA benchmark using the same scores as Table 2. Relative changes to Table 2 are shown in parentheses of the *Total Average* column. For the answerable questions, we report the F1 and Exact Match (EM) scores. For the unanswerable (unansrbl) questions, we only include the F1 score as the EM score is identical to it by definition. The average is taken across datapoints. The best score for each language in each column is bold.

		five-shot F1	five-shot EM
German	Creative Writing	52.89	43.07
	Product Reviews	54.77	45.27
	News	59.34	49.33
Turkish	Creative Writing	55.36	46.33
	Product Reviews	58.96	43.47
	News	55.72	41.67
Chinese	Creative Writing	37.80	37.13
	Product Reviews	40.18	39.67
	News	30.81	26.73
Average		49.54	41.41

Table 7: Five-shot language transfer with `gpt-3.5-turbo-0613`. The prompts contain examples from the target domain. The prompt for German is written in Chinese, for Turkish in German and for Chinese in Turkish.

		zero-shot F1	zero-shot EM	five-shot F1	five-shot EM
German	Creative Writing	55.49	46.07	58.11	46.67
	Product Reviews	54.88	44.47	52.01	39.33
	News	60.88	49.87	60.46	48.40
Turkish	Creative Writing	36.93	28.07	52.71	44.27
	Product Reviews	45.52	30.40	51.88	35.80
	News	43.59	28.93	54.54	39.13
Chinese	Creative Writing	44.38	44.33	47.98	47.93
	Product Reviews	41.65	41.60	41.45	41.40
	News	34.27	32.47	31.14	30.20
Average		46.40	38.47	50.03	41.4

Table 8: Domain transfer with `gpt-3.5-turbo-0613`: This table evaluates zero-shot and five-shot prompts written in the target language. The five-shot prompt for creative writing contains examples from M2QA news, the prompt for news from M2QA product reviews and the prompt for product reviews from M2QA creative writing.

these questions, we randomly sampled 50 questions to analyze the responses of all GPT-3.5 models we evaluated, i.e. the responses of the zero-shot and five-shot `gpt-3.5-turbo-0301` and `gpt-3.5-turbo-0613`. We found that in 72% of the cases, the question and answer are correctly annotated in the data, but the model either makes erroneous predictions (58%) or generates a correct answer instead of extracting it (14%). We further identified issues with inconsistent annotations (22%, i.e. 4.6% of all German data), questions with multiple plausible answers (4%), and the evaluation metric (2%). We provide some representative answers in Table 10. The full evaluation can be

Language	Annotators Kept	Annotators Rejected	Questions Checked
German	66	10	760
Turkish	32	12	440
Chinese	33	9	420

Table 9: Number of annotators we kept and how many we have rejected due to poor quality. For each annotator, we checked 10 questions. If at least two questions were of poor quality, i.e. did not follow our guidelines, the annotator got rejected. The last column shows how many of the accepted and rejected questions we checked in total for quality.

found in the M2QA GitHub repository.[15]

E Ablation Studies Detailed Results

We conducted two ablation studies to provide a deeper analysis of the baseline results on the M2QA benchmark. This section shows the detailed results of these ablation studies.

E.1 SQuAD Metric - Adaptation for Chinese

The SQuAD 2.0 metric uses whitespace tokenization. We modify the implementation to make the metric applicable to Chinese texts without whitespace tokenization by splitting the text into tokens using the off-the-shelf jieba tokenizer[16]. The results are in Table 11.

E.2 Adding Whitespaces to Chinese Text

We find that the XLM-R-based models only return answers surrounded by whitespace. As a result, for texts without whitespaces in Chinese, the model either marks the question as unanswerable or returns the whole context as the answer span. To explore the impact of this issue, we re-run the XLM-R$^{\text{Base}}$ setup but add whitespace between the words as determined by jieba. The results are presented in Table 12.

F Results on SQuADv2 and XQuAD

To show that our baselines and adapter-based setups do not only work on M2QA, we evaluated them also on SQuADv2 (Rajpurkar et al., 2018), and XQuAD (Artetxe et al., 2020b). Important to note is, that XQuAD only contains answerable questions. The results are presented in Table 13.

G Annotation Process

The number of annotators that were rejected vs. accepted during the annotation process and how many questions were checked in total is shown in Table 9.

H Data Annotation Platform

To be able to fulfil all of our requirements, we have developed our own annotation platform. The source code, including the tutorial, i.e. the instructions for the crowdworkers, is published in the same GitHub repository as the dataset[17]. We used GitHub Copilot[18] as AI assistance during coding. The crowdworkers first land on an overview page, then complete the tutorial and finally annotate data for M2QA. An annotation session consists of the tutorial and the annotation of 11 passages. If an annotator completed the tutorial in a previous session, it is optional, and they are assigned 12 passages. We assume that one annotation session results in a total of 1 hour of work. An evaluation after 65 annotation sessions showed that the crowdworkers took a median of 59 minutes. We pay crowdworkers £9 per annotation session, which is Prolifics recommended pay per hour. The tutorial consists of 3 steps in which the annotator is subsequently introduced to the task and learns to use the data annotation platform:

1. On the first page, they get an introduction to the annotation task.
2. Then they learn what makes good answerable questions and what to avoid when creating them.
3. Last, they learn what requirements good unanswerable questions must fulfil and what to avoid when creating them.

Figure 6 shows the interface that annotators use to annotate passages. Following SQuAD (Rajpurkar et al., 2016), we encourage annotators to pose hard questions in their own words. Since the wide adoption of LLM chatbots, the concern has arisen that crowdworkers could increasingly use LLMs to generate data instead of creating it themselves (Veselovsky et al., 2023). By disabling copy-pasting and requiring manual highlighting of the answer spans, we believe that using a ChatBot is not efficient in our setup. We found no evidence of the usage of LLMs during our quality checks.

[15] Link upon acceptance

[16] https://github.com/fxsjy/jieba v0.42.1

[17] Link upon acceptance

[18] https://github.com/features/copilot

ID	Question	Passage Text	Expected Answer	Answer by five-shot gpt-3.5-turbo-0613	Reason Why Answer Is Wrong
de_news_125-0_q0	Welche Position spielt Marc Janko?	Vorsichtig gab sich auch Stürmer Marc Janko: Uns Spielern ist die Schwere des Gegners bewusst, wir hatten in Moldawien ein sehr hektisches und schwieriges Spiel, erinnerte er an den knappen 2:1-Auswärtssieg im Oktober. Es war eine Partie, in der die österreichische Nationalelf mit der Spielweise Moldaus so manches Problem hatte. Am Samstag wird der moldauische Teamchef Alexandru Curtianu auf zwei Schlüsselspieler verzichten müssen: Der 28-jährige Abwehrstratege Alexandru Epureanu – mit 58 Einsätzen einer der erfahrensten Teamspieler – fällt wegen eines Kreuzbandrisses monatelang aus. Der Kapitän, der für Medipol Basaksehir, den Zwölften der türkischen Süperlig, aufläuft und früher zur Stammformation von Dinamo Moskau gehörte, ist mit einem Marktwert von 4,5 Millionen Euro der wertvollste Spieler der moldauischen Nationalelf.	Stürmer	unanswerable	Question and expected answer fine; model made a wrong prediction
de_books_2_61_q0	Was schlägt Klamm vor?	„Wollen wir es nun trotzdem versuchen, dennoch versuchen, ein Bündnis zu schließen? Wollen Sie meine Frau werden? Können Sie dem Vorurteil begegnen, daß ich nicht als der Freiherr von Klamm auftrete, der als Mann einer sehr reichen Frau lediglich die Zeit stiehlt und im Müßiggang lebt, sondern ein Geschäft, ein Gewerbe betreibt, arbeitet, schafft, fördert, maßvoll lebt, den rechten Lebensgewinn in dem Verkehr mit gleichgesinnten, wertvollen Personen erblickt, die denselben Anschauungen huldigen, so überlegen Sie meinen abermaligen Antrag! Aber gönnen Sie mir auch — verzeihen Sie das viele — das Gelöbnis, daß Sie lediglich mein sein und bleiben wollen, daß Sie“ — Klamm sprach's mit einem sanften, gewinnenden Lächeln — „keine anderen Götter haben wollen, neben mir!“	ein Bündnis zu schließen	Klamm suggests getting married.	Question and expected answer fine; model generated a correct answer instead of extracting it (often in english)
de_news_142-0_q2	Wer ist bald fuer NGOs zustaendig?	-H Chinas Parteibürokratie sieht das anders. Ihr Verbot scheint Teil jüngster Willkür-Maßnahmen in der reideologisierten Innenpolitik Chinas zu sein, um die Zivilgesellschaft unter ihre Kontrolle zu bringen. Die Behörden statuierten mit der Schließung der Frauenitiative, die auch von der Ford-Stiftung unterstützt wird, ein Exempel für alle zu eng mit dem Ausland verbundenen NGOs. Peking steht kurz vor Verabschiedung eines repressiven neuen Gesetzes für Auslands-NGOs. Betroffen sind Bürgerinitiativen, Stiftungen und Vereine. Sie sollen sich neu registrieren lassen und müssen ihre Arbeitspläne und Finanzen offenlegen. Künftig sollen sie der administrativen Kontrolle der Polizei unterstehen, statt wie bisher den Zivilämtern.	Sie sollen sich neu registrieren lassen	die Polizei	low-quality annotations
de_news_26-1_q0	Welcher Partei gehört Heiko Maas an?	Bundesjustizminister Heiko Maas hat den Handgranaten-Anschlag scharf verurteilt. Das Ausmaß der Gewalt ist erschreckend, erklärte der SPD-Politiker am Freitag in Berlin. Die Täter dürfen nicht ungestraft davonkommen. Sie müssen konsequent ermittelt und bestraft werden, forderte er. Die Zunahme der Angriffe auf Flüchtlinge sei dramatisch. Sprengkörper auf Flüchtlingsheime fliegen heute schon, wir dürfen nicht abwarten, bis es die ersten Toten gibt. Ähnlich äußerte sich der Zentralrat Deutscher Sinti und Roma. Dieser feige Anschlag zeigt, dass gewaltbereite Rechtsextremisten durch ihre Taten den Frieden in unserer Gesellschaft gefährden und uns auseinanderdividieren wollen, erklärte sein Vorsitzender Romani Rose. Umso mehr gelte es, für die Demokratie und den Rechtsstaat einzustehen. Besonders Politiker trügen hierbei eine große Verantwortung. Die populistische Rhetorik in der Asyldebatte führt dazu, dass Ängste bei der Bevölkerung geschürt werden, kritisierte Rose.	SPD-Politiker	SPD	problem with the evaluation metric
de_review_22_q2	Ist dieses Produkt empfehlenswert?	Ich bin begeistert. Dieses kleine Ding ist die Lösung auf all meinen Reisen. Wie oft ich mich geärgert habe, dass die sch*** Adapter nicht passen und ich lauter Netzgeräte einstecken musste, damit ich Handy, Kamera usw laden kann. Die Lösung kann so einfach sein. Absolut empfehlenswert. Zusätzliches Plus: Das Gerät besitzt eine eigene Sicherung (was in so manchem Ländern durchaus sinnvoll ist) und eine Ersatzsicherung wird gleich mitgeliefert. Würde auch 6* geben wenn ich könnte.	Absolut	empfehlenswert	multiple answers would be correct

Table 10: Samples of questions that five-shot `gpt-3.5-turbo-0613` failed at, along with the reason.

	Creative Writing F1 answerable	Product Reviews F1 answerable	News F1 answerable
XLM-R$^{\text{Base}}$	8.78 (+8.67)	8.82 (+8.13)	41.03 (+1.36)
XLM-R$^{\text{Domain}}$	7.24 (+7.24)	4.87 (+4.59)	1.94 (+0.15)
MAD-X+Domain	0.93 (+0.93)	2.73 (+2.34)	33.33 (+1.12)
MAD-X^2	4.32 (+4.21)	2.96 (+2.79)	33.57 (+0.33)
Llama 2-chat (13b)	18.52 (+5.47)	19.49 (+7.10)	28.97 (+18.11)
Llama 3-instruct (8b)	51.35 (+14.74)	43.07 (+15.55)	48.49 (+14.99)
gpt-3.5-turbo-0301	45.96 (+18.84)	40.21 (+20.71)	46.87 (+28.01)
gpt-3.5-turbo-0613	48.22 (+12.91)	40.54 (+14.53)	49.27 (+21.93)
Aya-23 (8b)	50.30 (+14.86)	40.64 (+13.9)	55.84 (+5.45)

Table 11: Chinese results using the adapted SQuAD 2.0 metric with word tokenization instead of whitespace tokenization, affecting F1 scores on answerable questions. Relative changes to Table 2 are shown in parentheses. LLMs use zero-shot prompts.

	Creative Writing			**Product Reviews**			**News**		
	answerable		unansrbl	answerable		unansrbl	answerable		unansrbl
	F1	EM	F1	F1	EM	F1	F1	EM	F1
original text	8.78	0.11	32.33	8.82	0.56	35.67	41.03	24.44	49.33
+ jieba whitespaces	25.24	16.89	70.00	22.11	12.89	60.00	28.36	7.44	50.50

Table 12: Results of XLM-R$^{\text{Base}}$ on the original texts and with added whitespace, evaluated with the adapted SQuAD metric using a word tokenizer instead of whitespace tokenization.

Model		SQuADv2 F1	SQuADv2 EM	XQuAD English F1	XQuAD English EM	XQuAD German F1	XQuAD German EM	XQuAD Turkish F1	XQuAD Turkish EM	XQuAD Chinese F1	XQuAD Chinese EM	M2QA Total Average F1	M2QA Total Average EM
XLM-R$^{\text{Base}}$	(0-shot)	74.72	70.96	71.86	62.18	53.62	40.59	48.73	36.97	42.30	35.13	37.73	31.59
XLM-R$^{\text{Domain}}$	(0-shot)	73.44	70.66	67.27	57.56	37.26	29.41	16.93	12.52	18.78	18.87	36.36	33.26
MAD-X+Domain	(0-shot)	75.50	72.32	69.34	59.92	51.24	39.66	46.57	35.21	41.72	34.20	41.42	37.68
MAD-X^2	(0-shot)	77.03	74.11	68.70	59.83	51.65	40.34	21.24	16.47	30.16	24.71	41.89	38.28
Llama 2-chat (13b)	(0-shot)	38.98	30.02	64.74	46.47	46.10	30.67	25.12	14.62	19.09	6.47	17.95	10.09
Llama 3-instruct (8b)	(0-shot)	57.05	52.18	77.86	64.03	66.93	50.92	57.76	39.92	53.69	47.14	44.54	29.59
gpt-3.5-turbo-0301	(0-shot)	67.34	59.56	76.50	60.00	68.35	47.65	58.50	35.13	41.29	35.71	47.08	37.50
gpt-3.5-turbo-0613	(0-shot)	72.92	68.50	78.20	66.39	68.16	52.35	61.65	43.78	60.95	58.15	53.11	45.00
Aya-23 (8b)	(0-shot)	77.49	74.87	78.94	70.84	69.04	56.55	63.73	49.75	62.08	58.07	51.61	44.16

Table 13: Results of the base models and adapter-based methods on SQuAD, XQuAD and M2QA.

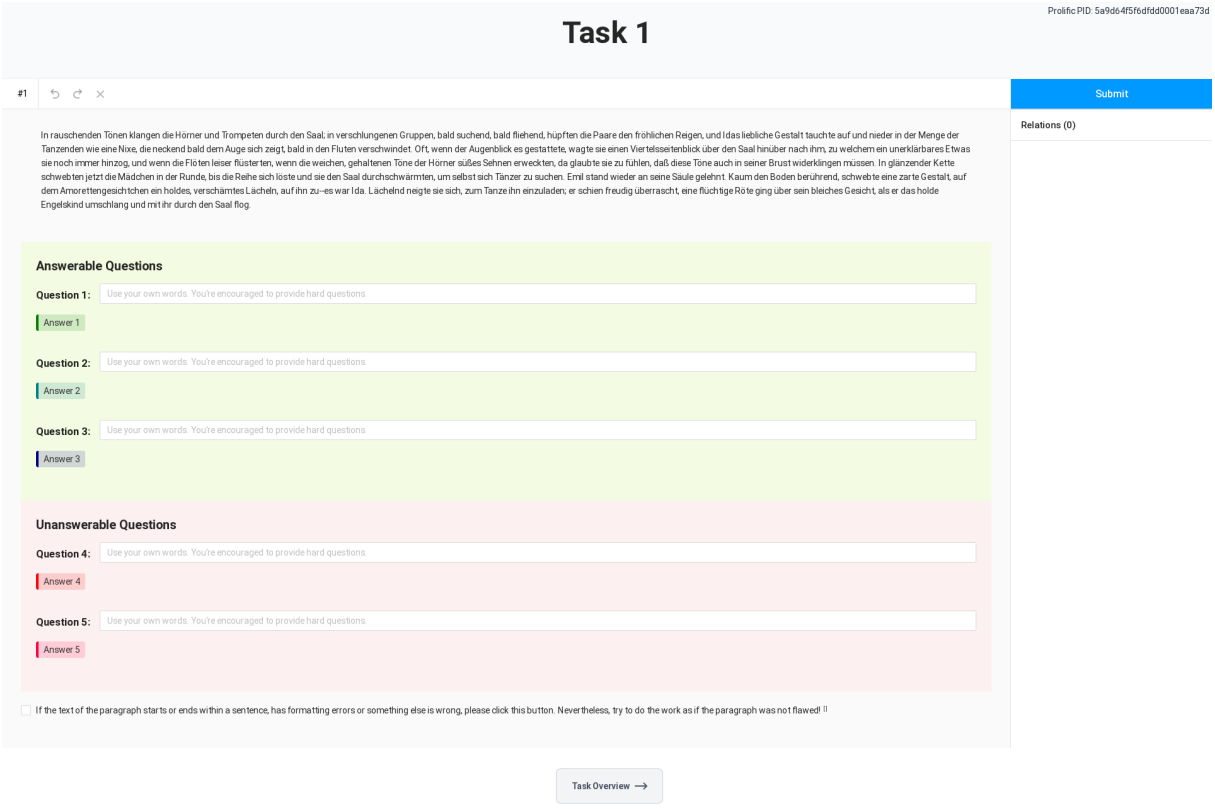


Figure 6: Screenshot of the interface that annotators use to write answerable and unanswerable questions and mark the respective answer span. Our interface is based on the Label Studio Frontend (Tkachenko et al., 2020).