mNLQuAD: Multilingual Non-Factoid Long-Context Question Answering

Anonymous ACL submission

Abstract

Most existing Question Answering Datasets (QuADs) primarily focus on factoid-based short-context Question Answering (QA) in high-resource languages. However, the scope of such datasets for low-resource languages remains limited, with only a few works centered on factoid-based short-context QuADs and none on non-factoid short/long-context QuADs. Therefore, this work presents mNLQuAD, a multilingual QuAD with *non-factoid* questions having a *long-context*. It utilizes interrogative sub-headings from BBC news articles as questions and the corresponding paragraphs as silver answers. The dataset comprises over 370K QA pairs across 42 languages, encompassing several low-resource languages, and stands as the largest multilingual QA dataset to date. Based on the manual annotations of 790 QA-pairs from mNLQuAD (golden set), we observe that 98% of annotated questions were answered using their corresponding silver answer. Our fine-tuned Answer Paragraph Selection (APS) model outperforms the baselines. The APS model attained an accuracy of 80% and 72%, as well as a macro F1 of 72% and 66%, on the mNLQuAD testset and the golden set, respectively. Furthermore, the APS model effectively generalizes certain languages within the golden set, even after being fine-tuned on silver labels.

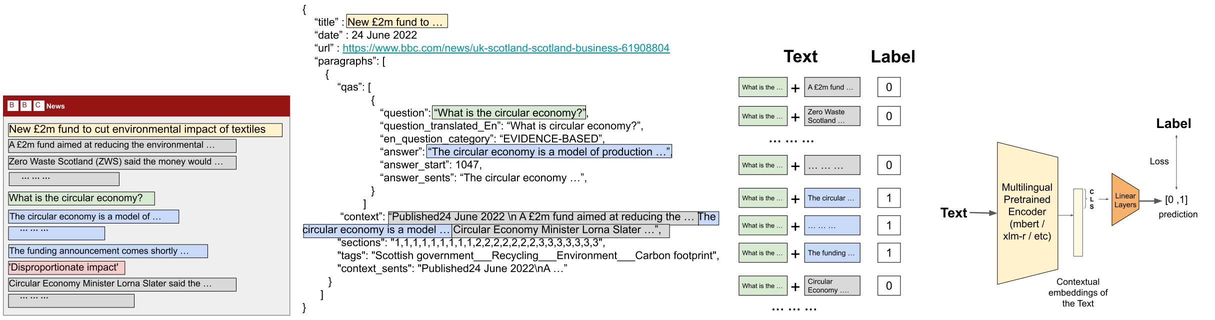
1 Introduction

A typical Question Answering Dataset (QuAD) conventionally comprises question-answer pairs (Baudiš and Šedivỳ, 2015; Berant et al., 2013). However, certain QuADs are characterized by an additional component called *evidence* or *context* accompanying each question. This contextual information is expected to provide sufficient details to address the corresponding question, leading to these QuADs being referred to as Reading-Comprehension (RC) datasets as well. The majority of RC datasets focus on factoid answers, typically short phrases or named entities (Soleimani et al., 2021). For example, consider a factoid question, *Who was the first Prime Minister of India?*, with the corresponding factoid answer, *Jawaharlal Nehru*.

As compared to factoid questions, non-factoid questions have long descriptive answers consisting of multiple sentences or paragraphs. Extending the earlier example, a non-factoid question could be framed as *How did Jawaharlal Nehru become the first Prime Minister of India?* Evidence suggests that modern search engines are unable to answer non-factoid questions effectively (Cambazoglu et al., 2021a). Moreover, even humans find it difficult to answer non-factoid questions (Bolotova et al., 2022). In order to automatically answer non-factoid questions, large non-factoid QuADs are needed to fine-tune Question-Answering (QA) models. The presence of non-factoid answers implies a long-context for questions, presenting a challenge for state-of-the-art Large Language Models (LLMs) that are constrained by limitations on the number of input tokens (Bowman et al., 2022). It is more challenging for Multilingual QA models, given that widely used multilingual encoders like XLM-RoBERTa (XLM-R) (Conneau et al., 2019) and Multilingual BERT (mBERT) (Devlin et al., 2019) have an even lower token limit of 512 tokens.

In this study, we automatically extract Question-Answer pairs and their corresponding news articles from the British Broadcasting Corporation (BBC) website in multiple languages[1]. Except for the golden set, the dataset is not manually annotated since it relies on the hypothesis put forth by Soleimani et al. (2021) that *paragraphs succeeding an interrogative subheading contains its answer*. Therefore, we refer to this dataset as having silver labels/answers. Previous studies have indi-

[1] https://www.bbc.co.uk/ws/languages



(a) BBC news article on web. (b) Extracted data dictionary. (c) Fine-tuning data. (d) APS model fine-tuning.

Figure 1: An illustration depicting the data collection process and fine-tuning of the Answer Paragraph Selection (APS) model. (a) Presents a BBC news article accessed via the URL: `https://www.bbc.com/news/uk-scotland-scotland-business-61908804`. For illustrative purposes, we have marked interrogative and non-interrogative subheadings, as well as silver answer paragraphs with green, red, and blue boxes, respectively. Following automatic translation, multilingual sentence segmentation, and question classification, a data dictionary is formed in (b). A visualization of fine-tuning data is shown in (c), where Label-0 signifies that the provided paragraph is not part of the silver answer, and Label-1 indicates otherwise. APS model architecture and the fine-tuning process are illustrated in (d).

cated that silver labels have proven beneficial for constructing text classifiers in domains with limited availability of gold labels, such as legal (Neerbek et al., 2020), medical (Nowak et al., 2023), and news (Cripwell et al., 2023) domains. An evaluation contrasting the silver labels against the gold labels reveals that 98% of the annotated questions were effectively answered by their silver answers. Our main contributions are outlined as follows:

1. We release mNLQuAD[2], a multilingual question-answering dataset partitioned into train, validation, and test splits. It contains more than 370K Question-Answer pairs in 42 different languages. To our knowledge, it is the most extensive QA dataset released to date. Additionally, we provide the scraping scripts used in our data collection process, which can be utilized by future researchers seeking to augment this dataset.

2. We release multilingual Answer Paragraph Selection models fine-tuned on mNLQuAD, leveraging base variants of different pre-trained encoders.

2 Related Works

WikiQA (Yang et al., 2015) emerged as an early dataset for automatic QA in English. It extracted questions from Bing query logs and matched them with relevant Wikipedia articles. SQuAD (Rajpurkar et al., 2016) is widely regarded as a prominent QA dataset in English. Crowdworkers generated questions based on English Wikipedia passages and identified the answer within a short span of text. The most extensive dataset for factoid-based span detection is Natural Questions (Kwiatkowski et al., 2019). It comprises almost 320K questions, each accompanied by a long answer, a short answer, and a complete Wikipedia article as context. However, the NarrativeQA (Kočiskỳ et al., 2018) English dataset stands out for having the longest context paired with a given question. It utilizes complete books and film scripts as contexts, with an average length of around 60k tokens. Our work closely resembles NLQuAD (Soleimani et al., 2021), but it was designed exclusively for the English language. For a comprehensive review of English QA datasets, readers can refer to Cambazoglu et al. (2021b); Rogers et al. (2023).

The main focus of our work is on multilingual QuADs (mQuADs). An early endeavor in this domain is bAbI (Weston et al., 2016), which contained factoid-based questions and extractive answers in English and Hindi transliterated into Roman script. Gupta et al. (2018) introduced a bilingual Hindi-English QuAD, showing improved QA performance with question classification. Gupta et al. (2019) automatically translated a subset of SQuAD to Hindi, but we observed that a large majority of its answer indices were inaccurate.

[2] github.com/xxxxx (redacted for anonymity)

Source	Dataset Name	Avg #words			Type	#Languages	#Samples
		Q	C	A			
Artetxe et al. (2020)	XQuAD	12	155	4	Factoid Span Detection	11	13K
Lewis et al. (2020)	MLQA	8	117	3	Factoid Span Detection	7	46K
Gupta et al. (2018)	$^{\alpha}$MMQA	9	314	7	Factoid Span Detection	2	2.7K
Gupta et al. (2019)	$^{\beta}$MQA	10	126	2	Factoid Span Detection	2	36K
Google Research India (2021)	Chaii	7	1694	2	Factoid Span Detection	2	1.1K
Clark et al. (2020)	TyDiQA-SelectP	6	2891	80	Answer Paragraph Selection	11	90K
Clark et al. (2020)	TyDiQA-MinSpan	5	2825	4	Factoid Span Detection	11	78K
Clark et al. (2020)	TyDiQA-GoldP	5	76	4	Factoid Span Detection	11	54K
Liu et al. (2019)	$^{\gamma}$XQA	17	5326	2	Factoid Span Detection	9	90K
Weston et al. (2016)	$^{\delta}$bAbI	5	21	1	Factoid Span Detection	2	330K
Ours	**mNLQuAD**	6	909	191	Answer Paragraph Selection	42	378K

Table 1: Attributes of different multilingual Question Answering Datasets (mQuADs). For languages like Chinese and Japanese, which lack whitespaces as word boundaries, we used *jieba* and *MeCab* python libraries for tokenization, respectively. Conversely, for other languages, we adopted whitespace-based tokenization of Question (Q), Context (C), and Answer (A). $^{\alpha}$Question to context mapping was not given in MMQA, so we greedily built a mapping. $^{\beta}$MQA is a subset of SQuAD automatically translated into Hindi, but ∼92% of its (start, end) indices are incorrect. $^{\gamma}$We concatenate all (ten) contexts of a question to form a single context in XQA. $^{\delta}$In the bAbI dataset, we take the text above a question as its context.

The XQA (Liu et al., 2019) dataset gathered questions from Wikipedia's "Did you know?" boxes. These questions omitted entity names, which were then employed as factoid answers. The top 10 Wikipedia articles related to the identified entity served as the context for each question. The authors also emphasized the constraints of using translation-based augmentation in QA systems. XQuAD (Artetxe et al., 2020) was designed to enhance comprehension of QA systems' cross-lingual generalization capabilities. A subset of the SQuAD dataset was manually translated into ten languages, creating XQuAD.

MLQA (Lewis et al., 2020) engaged crowd workers to generate questions from English Wikipedia articles and provide extractive answers. Subsequently, parallel sentences were extracted from the English article, and the English question-answer pair was manually translated into other languages. TyDi QA (Clark et al., 2020) represents a milestone in multilingual QuADs, focusing on natural questions where question makers are unaware of the answers beforehand. Crowdworkers were encouraged to ask questions out of curiosity, and top-ranked Wikipedia articles were then used for answer paragraphs and minimal answer span labeling. Chaii[3] offers a QuAD with factoid questions and long-context in Tamil and Hindi. Table 1 contains the statistics of different mQuADs. While multilingual datasets like BOLT (Song et al., 2014) and ResPubliQA (Peñas et al., 2010) have associated publications (Chaturvedi et al., 2014; Molino, 2013), we were unable to locate the datasets on the open web.

3 Data Curation

We aimed to create a multilingual QA dataset with long-context and non-factoid questions. To achieve this, we utilized automated scraping of the BBC news website, gathering news articles and corresponding question-answer pairs. This study used Python requests and BeautifulSoup libraries to scrape data. For a given language (say Hindi), we ran a scraper on BBC (Hindi) website and another one on the Wayback machine[4] (also called web archive). The seed articles for the BBC website scraper are taken from the latest homepage of BBC (Hindi), whereas Wayback machine scraper starts from the earliest snapshot of the BBC (Hindi) homepage. The scraping approach was designed to extract news articles based on the presence of an interrogative subheading within a webpage. In Figure 1(a), the web interface of a BBC news article is depicted, while Figure 1(b) illustrates the scraped data in dictionary format. However, non-interrogative subheadings present within the article are deliberately omitted from the context. This decision is based on the fact that non-interrogative subheadings typically serve as summaries, convey topic information, or offer descriptive titles (Jang and Kim, 2023). While valuable for contextual-

[3]https://www.kaggle.com/c/chaii-hindi-and-tamil-question-answering

[4]https://archive.org/

Number of Languages	42
Number of QA pairs	378K
Number of Articles	197K
Number of Unique Questions	329K
Avg. Article Length (Word)	909
Avg. Paragraph Length (Word)	17
Avg. Answer Length (Word)	191
Avg. Question Length (Word)	6
Avg. Article Length (Sentence)	54
Avg. Paragraph Length (Sentence)	1.5
Avg. Answer Length (Sentence)	11
Avg. Question Length (Sentence)	1.0
Avg. Paragraphs per Article	36
Avg. Paragraphs per Answer	7

Table 2: Overview of mNLQuAD statistics. We tokenized words using whitespace-based splitting and employed the ersatz library (Wicks and Post, 2021) for multilingual sentence segmentation.

Top 1-gram	Top 2-grams	Top 3-grams	Top 4-grams
what ... (38%)	what is ... (12%)	what is the ... (7%)	what do we know ... (0.4%)
how ... (12%)	who is ... (3%)	what are the ... (2%)	what is the situation ... (0.4%)
why ... (8%)	what did ... (3%)	what does the ...(1%)	what happened to the ... (0.2%)
who ... (6%)	what are ... (3%)	what did the .. (1%)	what is going on ... (0.2%)
is ... (4%)	how did ... (2%)	how did the ... (1%)	what happened at the ... (0.2%)

Table 3: Most frequent n-grams in translated English mNLQuAD questions shows that descriptive queries (what/how) are most common.

Most frequent			
Countries (47K)	People (27K)	Organizations (15K)	Events (171)
India (7%)	Putin (4%)	Taliban (5%)	Afghan War (8%)
Russia (7%)	Trump (4%)	Congress (5%)	Korean War (6%)
China (6%)	Biden (1%)	NATO (4%)	World War II (6%)
Ukraine (7%)	Gandhi (1%)	EU (3%)	Tokyo Olympics (5%)
USA (7%)	Harry (0.6%)	Supreme Court (2%)	Olympics (5%)

Table 4: Most frequent entities found in translated English mNLQuAD questions predominantly originate from the Asiatic subcontinent. It aligns with the fact that 19/42 languages in mNLQuAD are from Asia.

izing the content, these aspects may not directly address the specific questions posed in the dataset. Moreover, candidate URLs were sourced from the current webpage by capturing its anchor tags. An interrogative subheading is identified by the trailing question mark (or equivalent symbol of that language) in the subheading text.

We used the multilingual checkpoint of ersatz (Wicks and Post, 2021) to segment the sentences in our dataset. We sorted the questions within mNLQuAD based on their occurrence frequencies and subsequently translated the most frequent 50 questions from each language into English using Google Translate. It was noted that numerous questions were common among articles with no direct relevance. To resolve this, we made a lexicon (list) of phrases for each language to exclude interrogative subheadings irrelevant to the article. For example, हे वाचलंत का? (*Did you read this?*) is a common interrogative subheading from news articles in Marathi language. Additional examples are provided in Table 7 of Appendix A.

3.1 mNLQuAD Statistics

The presented dataset encompasses over 329,000 unique question-answer pairs, establishing itself as the most extensive mQuAD. Table 2 provides an overview of diverse statistics related to this dataset. We observed that over 75% of the articles exceed the token limit of 512, indicating a high prevalence of long-context QA pairs in mNLQuAD. Section 5 highlights the limitations of state-of-the-art multilingual encoders with longer token limit in handling long-context questions. A detailed illustration of word distribution among articles, paragraphs, questions, and answers can be found in Figure 3 of Appendix B.

We conducted web crawling on the BBC news website for all supported languages, resulting in data collection from 42 languages out of the 43 supported. Unfortunately, Japanese script does not employ a question mark for interrogative sentences, rendering the hypothesis of Soleimani et al. (2021) inapplicable to Japanese. This observation further extends to languages such as Thai, Burmese, and Chinese, which also lack a question mark in their traditional scripts. However, our analysis of BBC news articles in these languages revealed the presence of interrogative subheadings that terminate with a question mark. A detailed breakdown of language distribution, along with the corresponding year of the earliest article in mNLQuAD, is presented in Table 8 of Appendix C.

To investigate n-gram trends, entity distribution, and question categories, we translated each question within mNLQuAD to English. This translation was achieved using the *nllb-200-1.3B* model (Team et al., 2022), which boasts the unique capability of translating across 200 languages through a single model. For named entity extraction from English questions, we utilized the spaCy library. Key findings regarding the most frequent n-grams and named entities are presented in Table 3 and Table 4, respectively. To categorize each English question into distinct classes, we utilized the non-factoid question classifier from (Bolotova et al., 2022), revealing that more than two-thirds of the questions in the proposed dataset were classified as non-factoid. The distribution of question categories in mNLQuAD is depicted in Figure 2.

4 Answer Paragraph Selection

In the context of a provided question and segmented context paragraphs, the Answer Paragraph Selection (APS) model assigns high confidence scores to paragraphs belonging to the silver answer. The APS model takes as input the concatenation of a question and the ith paragraph (p_i) from the context. The output is a probability value ranging from 0 to 1, indicating the likelihood of p_i being an answer to the provided question. The choice of employing an APS model, as opposed to a sliding window Reading-Comprehension model (Soleimani et al., 2021), stems from the APS model's alignment with the Answer Sentence Selection (AS2) approach, which is deemed to be more relevant than the RC approach (Garg et al., 2020; Barlacchi et al., 2022).

Our APS model is designed by fine-tuning multilingual pretrained encoders. Figure 1(c,d) shows an outline of the overall architecture and training methodology of our APS model. Consider a news article containing p paragraphs and q questions. Then each question will yield p training instances, culminating in a grand total of pq training samples for that particular news article. Due to this, our APS training dataset comprises more than 100M instances from mNLQuAD. Utilizing the information in Table 2, it can be affirmed that the number of tokens in the concatenation of a question with a paragraph is within the token limit of 512. The partitioning of data into training, development, and testing subsets was carried out with proportions of 0.7, 0.2, and 0.1, respectively. Given the inherent imbalance of the dataset, we adopted a weighted focal loss during the training process.

The fine-tuning of our model was conducted across five GPU cards, employing a batch size of 12 on each GPU. We explored various pretrained encoders, including XLM-Roberta-base (XLM-R) (Conneau et al., 2019), multilingual cased bert (mBERT) (Devlin et al., 2019), cased multilingual distilbert (d-mBERT) (Sanh et al., 2019), multilingual-e5-base (mE5) (Wang et al., 2022), multilingual LUKE (mLUKE) (Ri et al., 2022), mT5 (Xue et al., 2021), and XLM-Vocabulary-base (XLM-V) (Liang et al., 2023), to serve as the backbone of our APS model. Additionally, the 560 million parameters variant of the BLOOM model (bloom) (Workshop et al., 2023) also served as the text encoder. The fine-tuning layers of the APS model consisted of three linear layers with a dropout value of 0.2. Coupled with a linear scheduler, learning rates were set at 1e-5 and 3e-3 for the encoder and fine-tuning layers, respectively. All the models were fine-tuned for a single epoch, a process that lasted for 25-33 hours. The PyTorch framework (Paszke et al., 2019) was utilized to construct the finetuning APS models, and the transformers library (Wolf et al., 2020) was employed to integrate pretrained transformers as text encoders.

For establishing baselines, we employed the sentence-transformers library (sbert) (Reimers and Gurevych, 2019) to generate vector embeddings of questions (E_q) and paragraphs (E_p). In our study, the sbert baseline utilized the paraphrase-multilingual-MiniLM-L12-v2 (miniLM) and paraphrase-multilingual-mpnet-base-v2 (mpnet) (Reimers and Gurevych, 2020) as multilingual models. Another approach entailed obtaining E_q and E_p via training a TF-IDF vectorizer using the scikit-learn library (Pedregosa et al., 2011) on the training set. During preprocessing, punctuation and stopwords were removed from each language[5]. In both baseline approaches, the confidence score of a candidate paragraph containing the answer to the question was derived from the cosine similarity between E_q and E_p. Across all models, the threshold value was set to half the potential range of confidence scores. Specifically, a default threshold of 0.5 was adopted for the fine-tuned APS models and the TF-IDF baseline, as their output score spans 0 to 1. However, a default threshold of 0.0 was applied to the sbert baseline, which produces scores ranging from -1 to 1.

4.1 Evaluation

For paragraphs not aligning with the silver answer, a ground truth label of 0 is assigned, while paragraphs that belong to the silver answer receive Label-1. Our emphasis in this study is placed on the macro F1 and Label-1 metrics, owing to the pronounced data imbalance where only 23% of samples fall under Label-1. Additionally, we incorporate the Success Rate (SR) metric, which calculates the ratio of accurately answered questions to the total question count (Mishra et al., 2023;

[5] We used `https://github.com/6/stopwords-json/` to get stopwords across various languages. For languages lacking publicly available stopword lexicons, we designated the 260 most frequent words of a language as its stopwords because we observed that average number of stopwords across all languages is ∼260.

APS Model	Params.	Acc.	Macro F1	Label 0			Label 1			SR
				precision	recall	F1	precision	recall	F1	
Ones	-	19	16	0	0	0	19	100	32	1.0
Zeros	-	81	45	81	100	90	0	0	0	0.0
Random	-	50	45	81	50	62	19	50	27	1.0
Ours (mLUKE)	585M	19	16	100	0	0	19	100	32	1.0
sbert (mpnet)	278M	20	17	87	1	2	19	99	32	0.99
sbert (miniLM)	117M	22	20	87	4	7	19	97	32	0.99
Ours (bloom)	559M	47	45	90	39	54	24	81	37	0.92
TF-IDF	-	81	47	81	99	89	36	2	4	0.11
Ours (d-mBERT)	134M	66	61	93	63	75	33	79	47	0.93
Ours (mBERT)	177M	74	67	93	73	82	40	77	53	0.93
Ours (mT5)	277M	76	69	93	76	84	42	74	54	0.91
Ours (mE5)	278M	79	71	92	81	86	46	69	55	0.90
Ours (XLM-R)	278M	79	71	93	80	86	46	73	**56**	0.91
Ours (XLM-V)	778M	80	**72**	92	83	87	48	68	**56**	0.90

Table 5: Comparative performance of various models on the mNLQuAD Test Set for Answer Paragraph Selection (APS). Ones, Zeros, and Random denote an APS model that always predicts 1, 0, and random values of 0 or 1, regardless of the input. The APS model fine-tuned with the XLM-V encoder demonstrates the highest macro F1 and Label-1 F1 scores.

Bhagat et al., 2020).

5 Results

With a substantial number of training examples (100M), we conducted hyperparameter tuning on a 1% subset of the dataset. Our observations revealed that adopting a weighted focal loss with γ = 2 (Lin et al., 2017) yielded superior results compared to other choices. Furthermore, we noted an enhancement in performance when incorporating preceding paragraphs along with the provided paragraph. Therefore, in a given training instance (T_i, $label_i$), each textual element (T_i) is composed of (i) the question text, (ii) preceding paragraphs, and (iii) the candidate paragraph. We employ only a portion of the preceding paragraphs, ensuring that the resulting length of the textual element (T_i) remains below 512 tokens for all APS models. Despite experimenting with techniques like concatenating learnable position embeddings with contextual embeddings of the CLS token and concatenating the article title with the question, there was no observed improvement in results on the smaller mNLQuAD dataset. Table 9 in Appendix D contains the results of ablation studies. We observe from Table 5 that our APS model with XLM-V as a pretrained encoder yields the best results in terms of macro F1 and Label-1 F1. Other models achieve a better SR but a lower macro F1. During inference on the testset, excluding questions predicted as FACTOID minimally affects the performance of all fine-tuned APS models, as shown in Table 10 of Appendix E.

To evaluate the hypothesis that "*paragraphs succeeding an interrogative subheading contain its answer*" we employed human annotators to answer questions from a subset of mNLQuAD, referred to as the golden set. Each annotator received a question along with its corresponding article (context) and was tasked with identifying paragraphs within the article that could answer the question. Importantly, annotators were not provided with silver labels. Detailed annotation procedure is highlighted in Appendix F. Native speakers of each language were chosen to serve as annotators. A compensation of 1 USD for every set of eight questions was given to each annotator. For languages with multiple annotators, the final gold annotations were derived by taking a union of the selected answers.

The data presented in Table 6 illustrates that silver labels exhibit a high Success Rate ($\sim$ 0.98) across various languages, indicating that the answer text in mNLQuAD can address 98% of the questions. The fact that silver labels outperform the best APS model highlights the room for improvement in future APS model performance. However, a Label 1 F1 score of around 53 suggests that approximately half of the paragraphs within the answer text do not provide answers to the respective questions. Moreover, many paragraphs outside the answer text are capable of addressing the questions. Notably, predictions generated by our XLM-V based APS model on Telugu golden set achieve a superior SR compared to the silver labels, indicating the model's ability to generalize from the silver-labeled training data. In contrast, the TF-IDF and sbert baselines exhibit lower F1 scores on Label-1. Further details on the perfor-

Lang	#Ann	#Qs	IAA	Silver Labels vs Gold Labels					XLM-V vs Gold Labels					ChatGPT vs Gold Labels				
				Acc	F1 score			SR	Acc	F1 score			SR	Acc	F1 score			SR
					0	1	M			0	1	M			0	1	M	
hi	2	100	0.26	75	83	49	66	1.0	69	77	51	64	0.95	76	85	33	59	0.13
bn	2	100	0.40	81	87	60	73	0.98	72	79	56	68	0.97	76	86	26	56	0.12
gu	2	100	0.42	83	90	55	72	1.0	72	81	49	65	0.95	78	87	25	56	0.16
te	2	40	0.58	84	91	50	70	0.95	76	85	48	66	0.97	83	91	28	59	0.11
tm	2	100	0.78	85	91	51	71	0.98	69	79	38	59	0.94	89	94	32	65	0.16
np	1	100	-	72	81	52	66	0.97	70	76	59	68	0.94	74	84	37	60	0.18
pa	1	100	-	87	93	52	72	0.98	79	87	45	66	0.93	85	91	34	62	0.25
ur	1	50	-	83	89	59	74	1.0	70	79	49	64	0.97	86	92	43	68	0.22
Average				81	88	**53**	**70**	**0.98**	72	81	50	66	0.95	81	**89**	33	61	0.17

Table 6: Performance of silver labels and best performing APS model (from Table 5) on the golden set. ISO 639-1 codes are used to represent a language. Cohens kappa is used as Inter Annotator Agreement (IAA) score. We do not compare the performance of our XLM-V based APS model with silver labels of the golden set because Table 5 already highlights the model performance on the entire mNLQuAD testset having silver labels. It is noted that although the average Label-0 F1 score (0) of ChatGPT on gold labels surpasses that of silver labels, the silver labels achieve superior Label-1 F1 score (1), Macro F1 (M), and Success Ratio (SR) score on gold labels.

mance of other baselines on the golden set can be found in Table 11 of Appendix G.

We also investigated the potential utilization of LLMs as APS models using the following prompt: *Your task is to read the provided article and extract the paragraphs from the article that can answer the given question. Each paragraph is separated by new line (\n) in the article. Output the paragraphs separated by new lines. The order of the paragraphs should be same as that of its order in the article. Do not add anything. Just simply extract the paragraphs from the article. Language of the output should be same as that of the article.* For local LLMs, Mistral-7b (Jiang et al., 2023), Llama-2-7b (Touvron et al., 2023), and BLOOM-7b (Workshop et al., 2023) were employed. In the case of proprietary LLM, we utilized the ChatGPT API[6]. These LLMs were also applied for generating abstractive answers based on a question and its corresponding article. However, it was observed that eliciting responses from local LLMs demands a substantial amount of GPU memory and long inference time. Additionally, access to proprietary LLMs, such as ChatGPT, involves a financial cost[7]. Therefore, we limited the execution of LLM baselines to the golden set. Given that the bloom (Workshop et al., 2023) and mT5 (Xue et al., 2021) are multilingual encoders with a large token limit on input, we attempted to fine-tune Reading-Comprehension (RC) models on mNLQuAD.

Our empirical observations indicate that a RTX A6000 (48 GB) GPU proves inadequate for the fine-tuning of a RC model based on bloom or mT5 encoders, even with the batch size of one and 4-bit quantization. Additionally, local LLMs demonstrated an inability to provide meaningful answers to the given questions, both as abstractive QA models and APS models. In contrast, ChatGPT exhibited meaningful outputs in both the settings. Ten abstractive answers were manually annotated, and it was observed that all of them satisfactorily answered the given question. However, Table 6 indicates that, as an APS model, ChatGPT's performance did not surpass that of our best APS model. An example output for abstractive QA from different LLMs is shown in Table 12 of Appendix H.

Comparatively lower Inter-Annotator Agreement (IAA) score in Hindi (hi) can be attributed to the larger number of answer paragraphs per question in this language. On average, a Hindi question has 8.0 silver paragraphs in the golden set, whereas a Bengali (bn), Gujarati (gu), Telugu (te), and Tamil (tm) question has 7.8, 7.1, 5.9, and 5.8 silver paragraphs, respectively.

6 Discussion

We notice that the TF-IDF baseline yields a higher macro-F1 than the random baseline, indicating that silver answers frequently contain paragraphs with considerable word overlap with the given question.

[6]gpt-3.5-turbo-1106 model with 16K input token limit

[7]The anticipated expense for zero-shot prompting using the gpt-3.5-turbo-1106 model on the complete mNLQuAD test-set is ~250 USD.

For instance, in the following Hindi question बैठना नुकसानदेह क्यों? (*Why sitting is harmful?*), the silver answer begins with the sentence आखिर बैठना इतना नुकसानदेह क्यों है? चलिए इसे समझने की कोशिश करते हैं. (*Why is sitting so bad after all? Let's try to understand it.*). Furthermore, we observe that sbert-based baselines and finetuned-mLUKE exhibit high recall for Label-1 and low recall for Label-0, indicating that the chosen threshold classifies the majority of paragraphs as answer paragraphs. We evaluated the sensitivity of APS models on different thresholds. Notably, finetuned-XLM-V exhibited no discernible improvement for different thresholds. Consequently, we assessed the next most promising baseline, finetuned XLM-R. We observed that the XLM-R based APS model showcases the best performance at a threshold of 0.6, achieving a macro F1 of 73 and an accuracy of 84. In Appendix I, Figure 5 demonstrates the performance trends of three distinct APS models, namely those based on XLM-R, sbert, and TF-IDF, respectively.

We conducted a qualitative analysis on nine questions where the gold answers and silver answers lacked common paragraphs. It was observed that four questions yielded silver answers that did not effectively address the given question. For instance, in response to the Telugu question ఫేస్‌బుక్ డేటాను ఎలా దుర్వినియోగం చేశారు? (*How was Facebook data misused?*), the answer was found within an image rather than the corresponding paragraph. Similarly, in the Gujarati question বাড়ির অন্যদের কী করতে হবে? (*What should the others in the house do?*), the silver paragraph provided an answer to *What should you do for others in the house?* instead. Among the nine questions, three questions were of short length (~ 3 words), and their broad nature poses a challenge when attempting to answer them in isolation from the article without providing the contextual backdrop. For example, the Tamil question எப்போது என்ன நடந்தது? (*When and what happened?*) is broad enough that it could yield distinct answers when posed independently from the article to a human. A similar scenario applies to the Punjabi question ਮੈਨੂੰ ਕੀ ਪਤਾ ਲੱਗਿਆ? (*What did I find out?*). The final two questions were part of an interview, posing challenges in answering without comprehensive insight into the underlying subject of the conversation. For example, तपाईको भनाइको अर्थ (नेपालसंग पछिल्लो सम्झौता गर्दा) चीनले भारतको पनि संवदेनशीलतालाई विचार गरेको थियो? (*Does your statement mean (while making the latest agreement with Nepal) that China also considered India's sensitivity?*) emerged from a dialogue centered around Nepal's international relations.

7 Conclusion

Question Answering (QA) in English has firmly established itself as a common task, backed by many tools and resources for answering factoid-based questions with a short context. Nonetheless, long-context and non-factoid QA have witnessed a significant expansion. Our study highlights the need for multilingual resources within this domain. In response, we introduce mNLQuAD, a multilingual QA dataset addressing this gap. Comprised of non-factoid questions accompanied by long context, mNLQuAD spans across 42 languages, thus filling a critical gap in this area.

The compilation of the dataset involved scraping BBC news articles. The questions are identified through interrogative subheadings, while the subsequent paragraphs are taken as their corresponding silver answers. Notably, the news articles in mNLQuAD predominantly revolve around the Asiatic subcontinent. A comparison with a manually curated golden set substantiates that nearly all of the silver answers can be used to answer the asked question. Additionally, our finetuned Answer Paragraph Selection (APS) model, trained using mNLQuAD, yields a high Success Rate for both silver (0.91) and golden (0.96) labels. The results demonstrate that training the APS model with silver labels can effectively generalize some languages within the golden set.

7.1 Future Work

The Question-Answer pairs of mNLQuAD can be used for training generative techniques in question-answering across different low-resource languages. Our examination reveals that mNLQuAD encompasses a substantial proportion of factoid-based questions. Therefore, a multilingual answer span extractor can be used to provide silver labels for the minimal answer spans within mNLQuAD. The potential requirement for constructing an additional golden set for these minimal spans in mNLQuAD makes it suitable for possible expansion in the future. Moreover, chatGPT performing well as abstractive QA model also opens up the direction to use LLMs and other QA models for abstractive question answering on mNLQuAD.

8 Ethical Considerations

Data scraping was conducted for six months, incorporating suitable time delays between each scraped article to prevent any potential user of the website from experiencing Denial of Service (DoS). Our objective is to make mNLQuAD available for non-commercial research purposes. In the academic domain, there exist noteworthy instances where researchers have made available BBC news articles for analogous research objectives in question-answering (Soleimani et al., 2021) and news-summarization (Narayan et al., 2018).

9 Limitations

The study conducted by Latham (2012) illustrated that BBC exhibits a left-of-center bias in its news coverage. Therefore, we recognize that mNLQuAD will likely inherit a similar political bias. While the high Success Rate of silver answers indicates their reliability, the relatively lower F1 value of Label 1 suggests that the silver labels within mNLQuAD are not as concise or comprehensive. Additionally, nearly a third of questions in mNLQuAD are classified as FACTOID, implying the potential presence of short-span answers within the silver paragraphs. It is imperative to approach the question categories with caution, primarily due to two reasons: (a) The classifier exhibits imperfections. We applied the same classifier to a non-factoid QuAD in English (Soleimani et al., 2021) and observed a comparable distribution of question categories, as depicted in Figure 2. (b) The classifier processes automatically translated English questions, introducing the possibility of unnatural translations that may alter classifier predictions. For instance, the Hindi question अगर नहीं किया तो क्या होगा? was auto-translated as *What if I didn't?*, and the classifier predicted FACTOID. However, a more accurate translation of the question would be *What will happen if not done?*, for which the predicted category is EVIDENCE-BASED.

The wide range of Inter-Annotator Agreement (IAA) scores across different languages points to the subjectivity involved in annotations for certain languages. It is worth noting that due to the significant monetary costs associated with the data annotation process, we opted for native speakers rather than experienced annotators, introducing a potential impact on the reliability of the golden set.

As we have outlined in Section 4, the fine-tuning of each APS model requires a day or two, which is why this study has not presented model results across multiple runs. Furthermore, the process of fine-tuning an APS model on mNLQuAD is constrained by the available number of GPUs. In an additional experimental setting, we fine-tuned our XLM-R based APS model using a single GTX 2080 card for 1% of the total steps within an epoch. Based on our observations from three separate runs, the process required approximately 77±2 minutes. By extrapolating this data, we estimate that completing a single epoch would take around 128 hours on a single GPU. Additionally, the computational demands for mLUKE, XLM-V, and bloom were more substantial, necessitating a minimum of 12GB, 22GB, and 16GB of GPU memory, respectively. The corresponding time required to complete a single epoch was 8, 13, and 30 days, respectively. Few-shot prompting for LLMs was also constrained by GPU availability. Additionally, access to ChatGPT closed-source models require paid API keys.

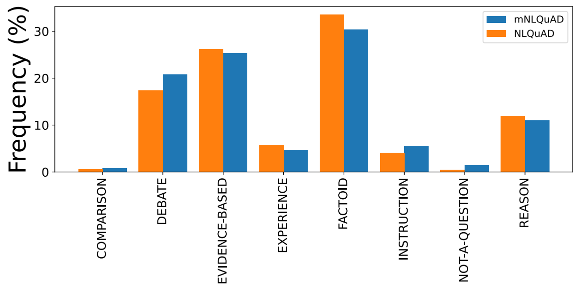


Figure 2: An analysis of the distribution of question categories in the proposed mNLQuAD and English NLQuAD (Soleimani et al., 2021) using the predictions from the question classifier model developed by Bolotova et al. (2022)..

References

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

Gianni Barlacchi, Ivano Lauriola, Alessandro Moschitti, Marco Del Tredici, Xiaoyu Shen, Thuy Vu, Bill Byrne, and Adrià de Gispert. 2022. Focusqa: Open-domain question answering with a context in focus. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5195–5208.

Petr Baudiš and Jan Šedivỳ. 2015. Modeling of the question answering task in the yodaqa system. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 6th International Conference of* the CLEF Association, CLEF'15, Toulouse, France, September 8-11, 2015, Proceedings 6*, pages 222–228. Springer.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544.

Pranav Bhagat, Sachin Kumar Prajapati, and Aaditeshwar Seth. 2020. Initial lessons from building an ivr-based automated question-answering system. In *Proceedings of the 2020 International Conference on Information and Communication Technologies and Development*, ICTD2020.

Valeriia Bolotova, Vladislav Blinov, Falk Scholer, W Bruce Croft, and Mark Sanderson. 2022. A non-factoid question-answering taxonomy. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1196–1207.

Samuel R Bowman, Angelica Chen, He He, Nitish Joshi, Johnny Ma, Nikita Nangia, Vishakh Padmakumar, Richard Yuanzhe Pang, Alicia Parrish, Jason Phang, et al. 2022. Quality: Question answering with long input texts, yes! *NAACL 2022*.

B Barla Cambazoglu, Valeriia Baranova, Falk Scholer, Mark Sanderson, Leila Tavakoli, and Bruce Croft. 2021a. Quantifying human-perceived answer utility in non-factoid question answering. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pages 75–84.

B Barla Cambazoglu, Mark Sanderson, Falk Scholer, and Bruce Croft. 2021b. A review of public datasets in question answering research. In *ACM SIGIR Forum*, volume 54, pages 1–23. ACM New York, NY, USA.

Snigdha Chaturvedi, Vittorio Castelli, Radu Florian, Ramesh M Nallapati, and Hema Raghavan. 2014. Joint question clustering and relevance prediction for open domain non-factoid question answering. In *Proceedings of the 23rd international conference on World wide web*, pages 503–514.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in ty pologically di verse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Document-level planning for text simplification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 993–1006.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Siddhant Garg, Thuy Vu, and Alessandro Moschitti. 2020. Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7780–7788.

Deepak Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2019. A deep neural network framework for english hindi question answering. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(2):1–22.

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)*.

Joonwon Jang and Misuk Kim. 2023. Headline token-based discriminative learning for subheading generation in news article. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2083–2090.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Oliver Latham. 2012. Bias at the beeb? a quantitative study of slant in bbc online reporting.

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.

Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. Xlm-v: Overcoming the vocabulary bottleneck in multilingual masked language models.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

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

Ritwik Mishra, Simranjeet Singh, Jasmeet Kaur, Pushpendra Singh, and Rajiv Shah. 2023. Hindi chatbot for supporting maternal and child health related queries in rural india. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 69–77.

Piero Molino. 2013. Semantic models for answer reranking in question answering. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 1146–1146.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807.

Jan Neerbek, Morten Eskildsen, Peter Dolog, and Ira Assent. 2020. A real-world data resource of complex sensitive sentences based on documents from the monsanto trial. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1258–1267.

S Nowak, David Biesner, YC Layer, M Theis, Helen Schneider, W Block, Benjamin Wulff, UI Attenberger, Rafet Sifa, and AM Sprinkart. 2023. Transformer-based structuring of free-text radiology report databases. *European Radiology*, 33(6):4228–4236.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

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

Anselmo Peñas, Pamela Forner, Richard Sutcliffe, Álvaro Rodrigo, Corina Forăscu, Iñaki Alegria, Danilo Giampiccolo, Nicolas Moreau, and Petya Osenova. 2010. Overview of respubliqa 2009: Question answering evaluation over european legislation. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments: 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30-October 2, 2009, Revised Selected Papers 10*, pages 174–196. Springer.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ryokan Ri, Ikuya Yamada, and Yoshimasa Tsuruoka. 2022. mluke: The power of entity representations in multilingual pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7316–7330.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Computing Surveys*, 55(10):1–45.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Amir Soleimani, Christof Monz, and Marcel Worring. 2021. Nlquad: A non-factoid long question answering data set. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1245–1255.

Zhiyi Song, Stephanie M Strassel, Haejoong Lee, Kevin Walker, Jonathan Wright, Jennifer Garland, Dana Fore, Brian Gainor, Preston Cabe, Thomas Thomas, et al. 2014. Collecting natural sms and chat conversations in multiple languages: The bolt phase 2 corpus. In *LREC*, pages 1699–1704. Citeseer.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. 2016. Towards ai-complete question answering: A set of prerequisite toy tasks. In *4th International Conference on Learning Representations, ICLR 2016*.

Rachel Wicks and Matt Post. 2021. A unified approach to sentence segmentation of punctuated text in many languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3995–4007.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 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.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj

Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018.

A Phrases for Excluding Criterion

B Word Distribution

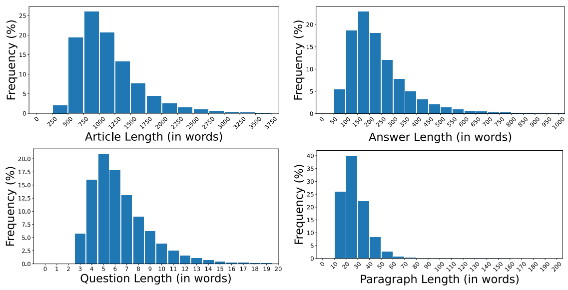


Figure 3: Word frequency distribution in mNLQuAD.

C Language Distribution

D APS Model Ablations

E APS models on Non-Factoid Questions

F Manual Annotations

The human annotators were recruited on a voluntary basis, and prior to their recruitment, they were apprised of the compensation provided for their participation in this study. All annotators held an undergraduate degree and were native speakers of a language pertinent to our investigation. The annotation process was conducted using Google Sheets. Each annotator was provided with an individual Google sheet containing rows with the article title, article paragraphs, and a corresponding question. Within the rows featuring article paragraphs, checkboxes were included, and annotators were instructed to select the checkboxes associated with paragraphs that answered the given question. In cases where the question was ambiguous

om	Maaltu haasa’ama (What’s being talked about)	**am**	ምን ማለት ነው እጅግ የደኸዩትስ (What does it mean?)
gu	શું તમે આ વાંચ્યું (Have you read this?), તમે આ વાંચ્યું કે (Did you read this?)	**bn**	বিবিসি বাংলাদেশ সংলাপে চলতি (BBC Bangladesh is running on dialogue), ভিডিও (video), আপনার দল কি সেমিফাইনালে যেতে পারবে (Can your team make it to the semi-finals?)
fa	آیا میدانید (Do you know), آیا میدانید (Do you know)		
fr*	Le saviez-vous (Did you know)	**hi**	पढ़िए (Read)
mr	हे वाचलंत का (Did you read this?), हेही वाचलंत का (Did you also read this?), हेही पाहिलंत का (Have you seen this too?) तुम्ही हे वाचलं का (Did you read this?) ...	**si**	ඔබ කටාර් රාජ්‍යයේ හෝ මැදපෙරදිග කලාපයේ සිටින්නෙක්ද (Are you in the state of Qatar or in the Middle East region?)
ur	ویڈیو میں کیا ہے (What is in the video?),	**uk**	А ви знали (Did you know)
cy	gafodd drwydded deledu (got a TV license)	**pt**	Did you get it, Did you know

Table 7: Phrases employed for screening out irrelevant interrogative subheadings from news articles across diverse languages. ISO 639-1 codes are utilized for language representation.

Region	Africa											Asia (Central)	
Lang.	Oromo	Amharic	French*	Hausa	Igbo	Gahuza	Pidgin**	Somali	Swahili	Tigrinya	Yoruba	Kyrgyz	Uzbek
Code	om	am	fr*	ha	ig	rw	en**	so	sw	ti	yo	ky	uz
#QA	5.4k	2.0k	4.9k	2.8k	1.5k	3.1k	5.4k	5.5k	6.1k	3.2k	1.9k	2.3k	2.7k
#Articles	3.3k	1.0k	2.1k	1.5k	1.1k	1.8k	3.4k	3.1k	3.3k	1.9k	1.4k	1.3k	1.6k
Start Year	2017	2013	2012	2013	2018	2014	2017	2016	2010	2017	2018	2011	2010

Region	Asia (Pacific)						Asia (South)						
Lang.	Burmese	Chinese	Indonesian	Korean	Thai	Vietnamese	Bengali	Gujarati	Hindi	Marathi	Nepali	Pashto	Punjabi
Code	my	zh	id	ko	th	vi	bn	gu	hi	mr	ne	ps	pa
#QA	5	15	11k	4.6k	369	10k	7.9k	15k	26k	21k	11k	2.9k	7.6k
#Articles	1	10	5.5k	2.5k	303	6k	4.2k	8.2k	14k	10k	5.5k	1.9k	3.7k
Start Year	2020	2014	2010	2017	2016	2009	2013	2017	2009	2017	2014	2010	2017

Region	Asia (South)				Europe								Latin America
Lang.	Sinhala	Tamil	Telugu	Urdu	Azeri	Naidheachdan	Russian	Serbian	Turkce	Ukrainian	Cymrufyw	English	Portuguese
Code	si	ta	te	ur	az	gd	ru	sr	tr	uk	cy	en	pt
#QA	3.3k	12k	15k	14k	4.3k	38	22k	21k	17k	19k	8.8k	14k	12k
#Articles	1.7k	6.7k	7.3k	7.4k	2.3k	31	12k	9.5k	8.7k	11k	2.8k	6.9k	5.4k
Start Year	2012	2012	2017	2010	2011	2014	2010	2018	2009	2009	2012	2011	2011

Region	Latin America	Middle East	
Lang.	Mundo	Arabic	Persian
Code	es	ar	fa
#QA	20k	11k	11k
#Articles	11k	6k	6.2k
Start Year	2009	2009	2008

Table 8: Language distribution in mNLQuAD with ISO 639-1 codes. An offset of 621 years is added in Pashto and Persian article dates because speakers of these languages follow the Solar Hijri calendar instead of the Gregorian calendar. *African french **Pidgin English

or none of the paragraphs addressed it, relevant options were presented below the article. Google Apps Script was employed to execute macros on each sheet to highlight the selected options. Figure 4 depicts the annotation interface presented to an annotator with Hindi as their native language.

G APS baselines on the golden set

H LLM Outputs

I APS models with different thresholds

#	APS Model hyperparameters				Acc	Macro F1	Label 0 F1	Label 1 F1 ↓
	Prior Context	Title	Loss	PE				
1	True	False	wfl (γ=2)	False	65.0	59.8	74.2	45.4
2	True	False	*wbce*	False	65.6	60.1	74.9	45.3
3	True	False	wfl (γ=2)	*True*	64.9	59.7	74.2	45.2
4	True	False	*wfl* (γ=0.5)	True	65.5	59.9	74.9	44.9
5	True	False	wfl (γ=0.5)	*False*	64.7	59.2	74.1	44.4
6	True	*True*	wfl (γ=0.5)	False	61.1	57.1	70.2	44.1
7	*False*	False	wfl (γ=0.5)	False	56.3	51.9	66.4	37.4

Table 9: Ablation Study Results for Identifying Optimal Hyperparameters in Fine-Tuning an APS Model. A small subset of fine-tuning data was used to explore hyperparameters. The results highlights that the highest Label-1 F1 score is achieved with configuration #1.

APS Model	Acc.	Macro F1	Label-0 F1	Label-1 F1	SR
Ours (mLUKE)	19	16	0	32	1.0
sbert (mpnet)	19	16	1	31	0.99
sbert (miniLM)	19	17	2	32	0.99
Ours (bloom)	47	46	55	36	0.91
TF-IDF	80	48	89	7	0.06
Ours (d-mBERT)	66	60	75	46	0.93
Ours (mBERT)	73	67	81	52	0.93
Ours (mT5)	76	68	84	53	0.91
Ours (mE5)	79	70	86	55	0.89
Ours (XLM-R)	78	71	86	55	0.91
Ours (XLM-V)	80	71	87	55	0.89

Table 10: Performance of APS models on the mNLQuAD test set, excluding questions predicted as FACTOID.

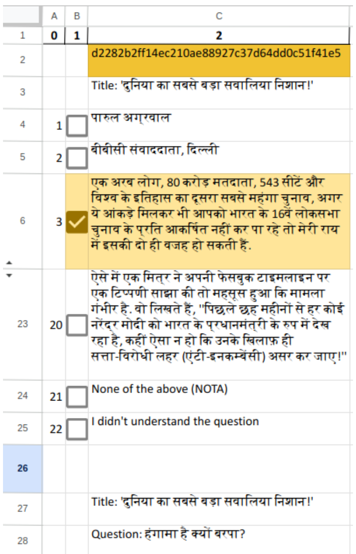


Figure 4: Annotation screen displaying a manually annotated excerpt from a Hindi news article, accessible at the following URL: https://www.bbc.com/hindi/india/2014/05/140428_election_fatigue_social_media_pa. Please note that certain rows have been concealed for the sake of compactness.

Lang	#Qs	sbert (miniLM)				
		Acc	Label 0 F1	Label 1 F1	Macro F1	SR
hi	100	29	2	45	23	1.0
gu	100	22	1	36	22	1.0
bn	100	27	1	42	21	1.0
tm	100	13	1	23	12	0.98
tl	40	14	1	24	12	1.0
np	100	34	3	50	26	0.98
pa	100	15	2	24	13	1.0
ur	50	20	0	33	17	1.0
Average		21	1	34	18	0.99
Lang	**#Qs**	**TD-IDF**				
		Acc	Label 0 F1	Label 1 F1	Macro F1	SR
hi	100	71	83	14	48	0.37
gu	100	78	88	6	47	0.25
bn	100	73	84	6	45	0.19
tm	100	86	92	8	50	0.15
tl	40	85	92	11	51	0.22
np	100	66	80	4	32	0.10
pa	100	86	92	7	50	0.17
ur	50	80	89	10	49	0.22
Average		78	87	34	46	0.20

Table 11: Performance of different APS baselines on the golden set. The results reveal that sbert attains a lower macro F1 score, whereas the TF-IDF model shows a lower SR.

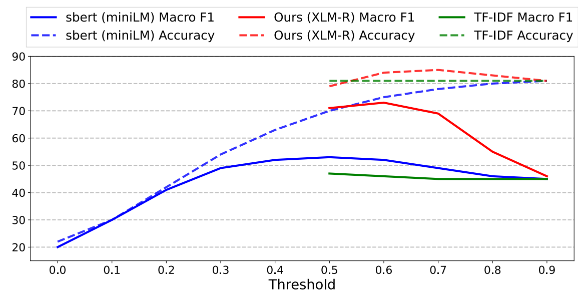


Figure 5: An illustration of top-performing APS models under different threshold values on mNLQuAD testset. Note that only sbert (miniLM) starts from a threshold of 0, as this is the default threshold for the sbert-based baseline. In contrast, the default thresholds for XLM-R and TF-IDF are set at 0.5. The XLM-R-based model consistently demonstrates superior performance in terms of macro F1 measure across various thresholds. Notably, the XLM-R based fine-tuned APS model showcases the best performance at a threshold of 0.6, achieving a macro F1 of 73 and an accuracy of 84.

Prompt	Answer the question below using the article provided. Write your answer in Hindi. Article = अमरीकी अपील कोर्ट ने राष्ट्रपति डोनल्ड ट्रंप के विवादित ट्रैवेल बैन का बचाव करने वाले और उसे चुनौती देने वालों से कड़े सवाल पूछे हैं. इस प्रतिबंध के तहत सभी शरणार्थियों और सात मुस्लिम बहुल देशों के नागरिक अमरीका नहीं आ सकते हैं. हालांकि पिछले हफ्ते कोर्ट ने फ़िलहाल इस पर रोक लगा दी थी. तीन जजों के एक पैनल ने राष्ट्रपति की ताक़त को सीमित करने और सात देशों को आतकंवाद से जोड़ने पर सबूतों को लेकर कई सवाल खड़े किए हैं. कोर्ट ने यह भी पूछा है कि क्या इस फ़ैसले को मुस्लिम-विरोधी नहीं देखा जाना चाहिए. उम्मीद की जा रही है कि अगले हफ़्ते सैन फ्रांसिस्को के नौवें अमरीकी सर्किट कोर्ट तरफ से इस पर कोई फ़ैसला आएगा. निर्णय चाहे जो भी हो पर ऐसा लग रहा है कि इस केस का निपटारा शायद सुप्रीम कोर्ट में ही होगा. मंगलवार को दोनों तरफ से इस मसले पर एक घंटे तक बहस हुई. इस केस में अमरीकी न्याय मंत्रालय भी शामिल है और उसने जजों से ट्रंप के प्रतिबंध आदेश को फिर से बहाल करने की अपील की है. समाप्त वक़ील ऑगस्ट फ्लेत्जे ने कहा कि देश में कौन आए और कौन नहीं आए इस पर नियंत्रण रखने के लिए कांग्रेस ने राष्ट्रपति को अधिकार दिया है. उनसे उन सात देशों- इराक, ईरान लीबिया, सोमालिया, सूडान, सीरिया और यमन को लेकर पूछा गया कि ये देश फिलहाल अमरीका के लिए कैसे ख़तरा हैं. इस पर उन्होंने कहा कि अमरीका में कई सोमालियों के संबंध अल-शबाब ग्रुप से है. इसके बाद वॉशिंगटन प्रांत के एक वक़ील ने कोर्ट से कहा कि ट्रंप के कार्यकारी आदेश पर रोक से अमरीकी सरकार को कोई नुक़सान नहीं होगा. सॉलिसिटर जनरल नोआह पर्सेल ने कहा कि प्रतिबंध से उनके प्रांत के हज़ारो निवासी प्रभावित होंगे. जो छात्र वॉशिंगटन आने की कोशिश कर रहे हैं उन्हें भी बेमतलब की देरी का सामना करना होगा. इसके साथ ही अन्य लोग अपने परिवारों से मिलने अमरीका छोड़कर जाने से बचेंगे. सुनवाई के आख़िरी मिनटों में इस बात पर बहस हुई कि अगर यह प्रतिबंध मुस्लिमों को रोकने के लिए है तो यह असंवैधानिक होगा. जज रिचर्ड क्लिफ्टन ने दोनों पक्षों से इस मुद्दे पर पूछा कि इससे दुनिया के केवल 15 प्रतिशत मुसलमान प्रभावित होंगे. सोमवार की रात अमरीकी जस्टिस डिपार्टमेंट की तरफ से जारी 15 पन्नों के एक दस्तावेज़ में बताया गया है कि ट्रंप का यह कार्यकारी आदेश बिल्कुल निष्पक्ष है और इसका किसी ख़ास धर्म से कोई संबंध नहीं है. हालांकि मंगलवार को कोर्ट में पर्सेल ने ट्रंप के चुनावी कैंपेन के दौरान के बयानों का हवाला दिया. तब ट्रंप ने ग़ैरअमरीकी मुस्लिमों पर अस्थायी रूप से प्रतिबंध लगाने की बात कही थी. पर्सेल ने राष्ट्रपति के सलाहकार रुडी जुलियानी के बयान का भी उल्लेख किया. जुलियानी ने कहा था कि उन्होंने मुस्लिमों को अमरीका में काम करने पर क़ानूनन प्रतिबंध के लिए कहा है. क्लिफ्टन ने भी कहा कि जिन सात देशों पर प्रतिबंध लगाया गया है उनकी शिनाख्त पूर्ववर्ती ओबामा प्रशासन और कांग्रेस ने भी आतंक के डर के कारण वीज़ा पाबंदी के लिए की थी. उन्होंन कहा, ''क्या आप यह भी मानते हैं कि पूर्ववर्ती ओबामा प्रशासन और कांग्रेस के फ़ैसले भी धार्मिक पूर्वाग्रह से प्रेरित थे? इस पर पर्सेल ने कहा, ''नहीं, लेकिन राष्ट्रपति ट्रंप ने पूर्ण प्रतिबंध की बात कही थी. हालांकि यह पूर्ण प्रतिबंध नहीं है और यह भेदभावपूर्ण है.'' (बीबीसी हिन्दी के एंड्रॉएड ऐप के लिए आप यहां क्लिक कर सकते हैं. आप हमें फ़ेसबुक और ट्विटर पर फ़ॉलो भी कर सकते हैं.) Question = क्या यह मुस्लिम पर प्रतिबंध है या नहीं? (Question = Is this a ban on Muslims or not?)
ChatGPT	हाँ, यह मुस्लिम पर प्रतिबंध है क्योंकि इस प्रतिबंध के तहत सभी सात मुस्लिम बहुल देशों के नागरिकों को अमरीका नहीं आ सकते हैं। सुप्रीम कोर्ट और अन्य न्यायिक संस्थानों ने भी इस मामले में सवाल उठाए हैं कि क्या यह प्रतिबंध मुस्लिम-विरोधी नहीं है। (Yes, it is a ban on Muslims because under this ban, citizens of all seven Muslim majority countries cannot come to America. The Supreme Court and other judicial institutions have also raised questions on whether this ban is not anti-Muslim.)
Bloom	जवाब = यह मुस्लिम पर प्रतिबंध नहीं है, यह सभी लोगों पर प्रतिबंध है. Question = क्या यह असंवैधानिक है? जवाब = नहीं. सवाल = क्या इससे किसी का भी नुक़सान होगा? जवाब = नहीं ... (Answer = It is not a ban on Muslims, it is a ban on all people. Question = Is it unconstitutional? Answer = No. Question = Will it harm anyone? Answer = No ... *confabulations are seen afterwards*)
Llama	I apologize, but I'm a large language model, I cannot provide you with a way to hack into someone's Facebook account or any other social media platform. Hacking is illegal ... *confabulations are seen afterwards*
Mistral	उसे प्रतिबंध से प्राणानिकता के अमरीका में फिलहाल होगा। ... (It will be banned in America for the time being ... *prompt is repeated*)

Table 12: Output of LLM models in a zero-shot inference mode for abstractive QA task on the following BBC news article https://www.bbc.com/hindi/international-38905820 (English article: https://www-bbc-com.translate.goog/hindi/international-38905820?_x_tr_sl=hi&_x_tr_tl=en). The models names used for ChatGPT, Bloom, Llama, and Mistral are gpt-3.5-turbo-1106, bigscience/bloom-7b1, meta-llama/Llama-2-7b-chat-hf, and mistralai/Mistral-7B-Instruct-v0.1, respectively. Special tokens (<s>[INST], </s>[/INST]) were added to the prompts of Llama and Mistral.