NativQA: Multilingual Culturally-Aligned Natural Query for LLMs

Anonymous ACL submission

Abstract

Natural Question Answering (QA) datasets play a crucial role in developing and evaluating the capabilities of large language models (LLMs), ensuring their effective usage in real-world applications. Despite the numerous QA datasets that have been developed, there is a notable lack of region-specific datasets generated by native users in their own languages. This gap hinders the effective benchmarking of LLMs for regional and cultural specificities. In this study, we propose a scalable framework, *NativQA*, to seamlessly construct culturally and regionally aligned QA datasets in native languages, for LLM evaluation and tuning. Moreover, to demonstrate the efficacy of the proposed framework, we designed a multilingual natural QA dataset, Multi*NativQA*, consisting of ∼72K QA pairs in seven languages, ranging from high to extremly low resource, based on queries from native speakers covering 18 topics. We benchmark the Multi*NativQA* dataset with open- and closed-source LLMs. We made both the framework *NativQA* and Multi*NativQA* dataset publicly available for the community.[1]

1 Introduction

Recent advancements in Large Language Models (LLMs) have revolutionized the landscape of artificial intelligence, significantly pushing the state-of-the-art for a broad array of Natural Language Processing (NLP) and Speech Processing tasks, such as machine translation, question answering, automatic speech recognition, text-to-speech generation among others. Their potential in language understanding and generation, across multiple (high- and low-resourced) languages has attracted researchers to benchmark the LLM capabilities across diverse tasks, domains, and disciplines (OpenAI, 2023; Touvron et al., 2023a,a). Moreover, the rapid integration of LLMs with various applications necessitates measuring cultural discrepancies in the responses generated by LLMs to ensure alignment with users' cultural values and contexts. Evaluating the generalization capabilities of LLMs across different tasks and languages has recently garnered significant attention. The HELM project (Liang et al., 2022) assessed English LLMs across various metrics and scenarios. BIG-Bench (Srivastava et al., 2023) introduced a large-scale evaluation with 214 tasks, including low-resource languages. Recently, GPT-2.5 (Radford et al., 2019), ChatGPT (OpenAI, 2023), and BLOOM (Scao et al., 2022) were evaluated by Bang et al. (2023); Ahuja et al. (2023); Hendy et al. (2023).

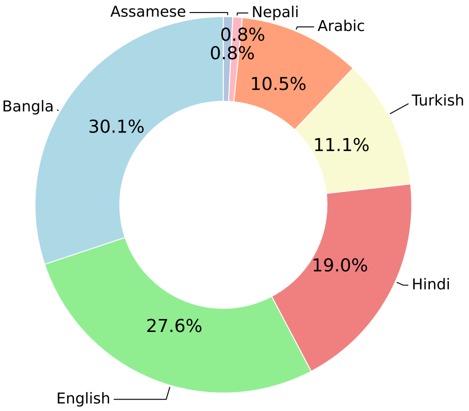


Figure 1: Distribution of the Multi*NativQA* dataset across different languages.

Such evaluations have been conducted on standard QA, NLP, and/or speech datasets. LLM developers measure different capabilities of their released models, such as *common sense reasoning* (e.g., HellaSwag (Zellers et al., 2019)), *world knowledge* (e.g., MMLU (Hendrycks et al., 2020) and Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017)), and *reading comprehension* (e.g., SQuAD (Rajpurkar et al., 2016)).

Interestingly, most of these datasets are in English. Hence, multilingual and non-English LLMs

[1] anonymous.com

Lang	Q/A	Example (Native)	English Translation
Arabic	Q	كم مساحة قطر طول وعرض؟	What is the area of Qatar length and width?
	A	يبلغ عرض مساحتها حوالي 100 كم وتمتد بطول 200 كم في الخليج.	Its area is about 100 km in width and extends 200 km in the Gulf.
Assamese	Q	কোন জন বিখ্যাত ৰাজনৈতিক ব্যক্তিয়ে শেহতীয়াকৈ অসমত বিজেপিৰ পৰা কংগ্ৰেছলৈ যোগদান কৰিছিল ?	Which famous political person recently joined from BJP to Congress in Assam?
	A	আমিনুল হক লস্কৰে শেহতীয়াকৈ অসমত বিজেপিৰ পৰা কংগ্ৰেছত যোগদান কৰিছিল।	Aminul Haque Laskar recently joined Congress from BJP in Assam.
Bangla	Q	শোলাকিয়া মাঠের আয়তন কত ?	What is the area of Sholakia field?
	A	বর্তমান শোলাকিয়া ঈদগাহ্ মাঠের আয়তন ৭ একর।	The current area of Sholakia Eidgah field is 7 acres.
English	Q	Does UDST offer scholarships?	NA
	A	Public schools in Qatar receive government funding and provide free tuition to all citizens.	NA
Hindi	Q	नवरात्रि में कलश रखने का शुभ मुहूर्त क्या है?	What is the auspicious time to keep Kalash in Navratri?
	A	कलश की स्थापना चैत्र शुक्ल पक्ष की प्रतिपदा तिथि को की जाती है. इस बार चैत्र नवरात्रि की घटस्थापना का सबसे अच्छा मुहूर्त सुबह 6 बजकर 2 मिनट लेकर सुबह 10 बजकर 15 मिनट तक है |	The Kalash is established on the Pratipada date of Chaitra Shukla Paksha. This time the best time for Chaitra Navratri is from 6.02 am to 10.15 am.
Nepali	Q	नेपालको सबैभन्दा ठूलो ताल कुन हो	Which is the biggest lake in Nepal?
	A	नेपालको सबैभन्दा ठूलो ताल कर्णाली प्रदेशको रारा ताल हो।	The largest lake in Nepal is Rara Lake in Karnali Province.
Turkish	Q	Istanbul'da göl var mı?	Is there any lake in Istanbul?
	A	İstanbul'da dört doğal göl bulunmaktadır. Bunların yanı sıra, baraj gölleri de vardır.	There are four natural lakes in Istanbul. In addition, there are also reservoir lakes.

Figure 2: Examples of questions and answers in different languages.

have been evaluated by using machine translation, with or without human involvement, to translate the existing English datasets into corresponding languages. For example, to evaluate Jais (Sengupta et al., 2023) and AceGPT (Huang et al., 2023), evaluation datasets have been translated into Arabic. Other examples on translated datasets include Korean MMLU (Son et al., 2024) and Okapi (Lai et al., 2023b), a translation of three benchmark datasets in 26 languages. Machine translation has its drawbacks in terms of accuracy, but most notably, inability to reflect cultural and regional specificities of a targeted language.

Consequently, we believe that using automatically-translated datasets is not ideal for an evaluation that truly reflects real users needs and tasks, and captures their cultural and regional interests and preferences as conveyed through their native languages. At the same time, the typical alternative of developing datasets in new languages by human annotators is a costly and time-consuming process.

Therefore, we propose a framework, *NativQA*, specifically designed to seamlessly develop regionally and culturally specific datasets following a human-machine collaborative approach. Datasets developed through *NativQA* serve two primary functions: *(i)* evaluating the LLM performance over real users information needs and interests expressed in their native languages, and *(ii)* facilitating fine-tuning and instruction tuning of LLMs to adapt to cultural contexts. Moreover, to show the efficacy of the *NativQA* framework, we developed a natural question-answering (QA) dataset, Multi*NativQA*, including ∼72K QA pairs in seven low to high resource languages (as shown in Figure 1), covering 18 different topics from various regions (see examples in Figure 2).

Our contribution in this study is as follows:

- We propose the *NativQA* framework for developing culture- and region-specific natural question-answering datasets. This framework helps enhance LLM inclusivity and provides comprehensive culturally-aligned benchmark datasets.
- We develop and release a dataset, Multi*NativQA*, in seven languages with a size of over 72K QA pairs, covering 18 different topics based on real queries from native speakers.
- We establish baselines and provide evaluation results over the Multi*NativQA* dataset using different open and closed models, promoting research in this area.

2 NativQA Framework

Figure 3 presents the *NativQA* framework consisting of three inter-connected modules – Query Collection, QA Collection and QA Validation.

2.1 Query Collection (QC)

The objective of this module is to gather open-ended queries focusing on various predetermined topics derived from common concepts in everyday communication. We believe that the set of topics should be manually constructed, as this step requires identifying topics that are culture- or region-dependent (e.g., Events, Literature, etc.) (see Table

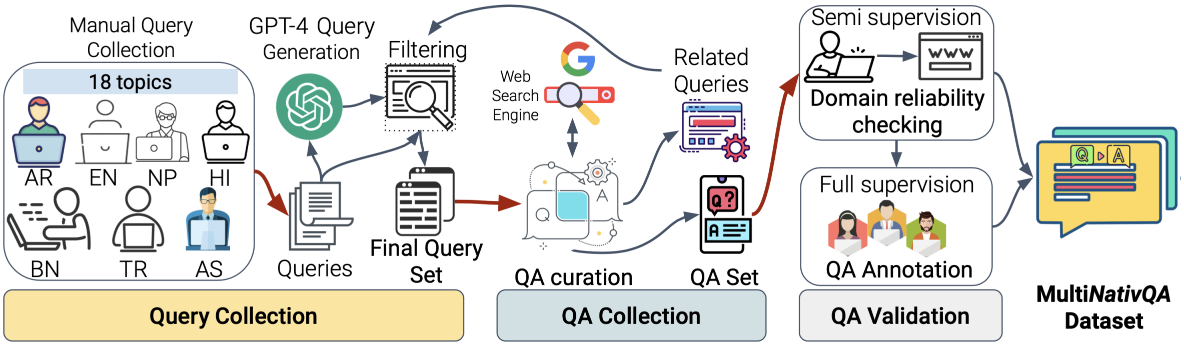


Figure 3: *NativQA* framework, demonstrating the data collection and annotation process.

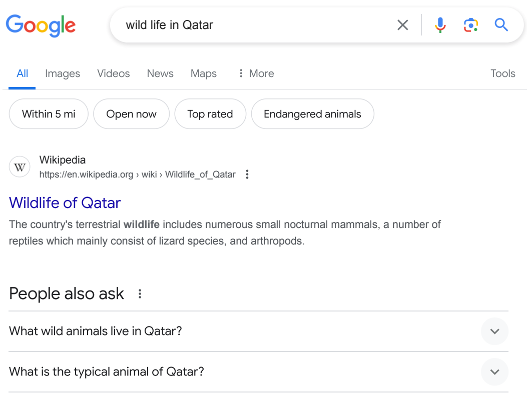
2). Next, query collection can start by recruiting native speakers in the target countries. Each speaker is encouraged to write M queries per topic, in their native (or second) language, focusing on issues they might encounter as residents of a corresponding major city.

Once the initial collection of queries, Q_m, is completed, the next step is to expand the set with synthesized queries, Q_s. The aim of expanding the set is to increase variability in sub-topics and writing styles in the final set of queries. For Q_s, we opt to prompt LLM to generate $x = 10$ similar queries for each input query, $q_m^i \in Q_m$. Finally, we de-duplicate Q_s against Q_m using exact string matching, resulting in the *final set* of seed queries, $Q_0 = Q_m \bigcup Q_s$.

2.2 QA Collection (QAC)

The next step is to collect QA pairs that potentially cover topics represented by Q_0, using a search engine, e.g., Google. We specifically select Google, since when a query is issued against it, it can return a data structure called “People also ask" that lists few questions asked by real users and potentially relevant to the initial user query, as shown in Figure 4. Moreover, the questions are associated with answers extracted by the search engine and links to the answers sources.

Our QA curation module implements Algorithm 1, using the seed queries Q_0 along with the number of iteration, N_{iter}, as input. For each iteration $i \in N_{iter}$, we collect QA pairs P_{QA}^i, and related queries S_{rel}^i for each query, $q \in Q$, and then pass it to the filtering module and update the current query set Q. We repeat the process for all the iterations to obtain the final QA set, S_{QA} for enriched queries Q.



Google

wild life in Qatar

All · Images · Videos · News · Maps · ⋮ More · Tools

Within 5 mi · Open now · Top rated · Endangered animals

Wikipedia
https://en.wikipedia.org › wiki › Wildlife_of_Qatar ⋮

Wildlife of Qatar

The country's terrestrial **wildlife** includes numerous small nocturnal mammals, a number of reptiles which mainly consist of lizard species, and arthropods.

People also ask ⋮

What wild animals live in Qatar?

What is the typical animal of Qatar?

Figure 4: Google's QA list in response to a query.

2.3 QA Validation (QAV)

The last step of the *NativQA* framework is to validate the QA pairs, considering at least two aspects: (i) the quality and answerability of questions, and (ii) reliability and completeness of answers. We validate the QA pairs through the following steps.

Domain Reliability Checking (DRC). The answers collected by our approach include a link to the web page from which an answer was extracted. Thus, our answer validation step starts by a semi-supervised approach that aims to keep QA pairs based on the reliability of the Web domain where the answer appears. We hypothesize that answers from web pages of a reliable domain are likely to be trustworthy. In our approach, unique domains for the QA pairs in S_{QA} are listed. Next, annotators manually annotate each domain by reliability based on an annotation guideline we designed for this task, inspired by several relevant studies (Selejan et al., 2016; Flanagin and Metzger, 2007; Metzger and Flanagin, 2015). We then only keep the QA pairs with answers from reliable sources.

For developing large-scale fine- and instruction-tuning data, this approach is practical and scalable

Algorithm 1 Collecting QA pairs using seed queries Q_0. P^i_{QA}: QA pair, S^i_{rel}: related queries. QA(*) and RQ (*) are functions that return questions and answers, and related queries, respectively, which are obtained from the search engine for a given query.

```
Input:
    Seed queries: $Q_0 = \{q_1, q_2, \ldots, q_n\}$
    Number of iterations: $N_{iter}$
Output:
    Set of QA pairs: $S_{QA}$
    Set of queries: $Q$
$S_{QA} \leftarrow \emptyset$
$Q \leftarrow Q_0$
for $i$ from 1 to $N_{iter}$ do
    $P^i_{QA} \leftarrow \emptyset$
    $S^i_{rel} \leftarrow \emptyset$
    for $q \in Q$ do
        $(Q^q, A^q) \leftarrow$ QA$(q)$
        $P^i_{QA} \leftarrow P^i_{QA} \cup \{(q', a') \mid q' \in Q^q, a' \in A^q\}$
        $S^i_{rel} \leftarrow S^i_{rel} \cup$ RQ$(q)$
    end for
    $P^i_{QA} \leftarrow$ filter_duplicates$(P^i_{QA})$
    $S_{QA} \leftarrow S_{QA} \cup P^i_{QA}$
    $Q \leftarrow Q^q_{rel} \cup S^i_{rel}$
end for
return $S_{QA}$
```

because it reduces manual effort required to obtain reliable answers. However, even if a domain is considered reliable (e.g., BBC, Guardian), this does not guarantee that the search engine effectively extracted an answer that accurately satisfies a given question. Thus, we added two more steps to validated the selected QA pairs as discussed below.

QA Annotation (QAA). Given the curated questions with answers from reliable sources, in this step, we opt for the manual checking and editing of answers for the remaining QA pairs. For each QA pair, we apply three types of annotations. *(i)* Question validation: human annotators should verify questions quality, and filter out lower-quality questions. Specifically, we define a "good question" as one that is fact-seeking, and can be answered with an entity or explanation. While a "bad question" is either ambiguous or incomprehensible, depends on clear false presupposition, opinion-seeking, or does not seek factual information. *(ii)* Answer categorization: for a good question, annotators are asked to categorize answers based on correctness (see Section 3.2.2) by examining each QA pair and assessing whether the answer provides sufficient information to satisfy the question. *(iii)* Answer editing: If an answer is not fully answering a question, annotator must edit the answer by adding more content from the answer source Web page such that the answer is complete. We limit the annotators to using the provided source Web pages to maintain the scope and reliability of answers we collect during this phase.

Lang.	Cat.	City	CC	#SQ	#QA	F.QA
Arabic	M	Doha	QA	3,664	12,311	7,548
Assamese	X	Assam	IN	900	21,009	572
Bangla	L	Dhaka	BD	889	13,688	10,724
Bangla	L	Kolkata	IN	900	13,378	10,969
English	H	Dhaka	BD	1,339	17,744	7,075
English	H	Doha	QA	3,414	25,621	12,806
Hindi	M	Delhi	IN	1,184	16,328	13,720
Nepali	L	Kathmandu	NP	1,222	11,503	581
Turkish	M	Istanbul	TR	900	23,143	8,027
Total		–	–	**14,412**	**154,725**	**72,022**

Table 1: List of languages with initial seed queries along with the number of QA pairs collected per language from different locations. CC: Country code, Lang.: Language, SQ: Seed Query. F. QA: Final QA set. Cat.: Categorization in terms of high (H), medium (M), low (L), and extremely low (X) as per (Lai et al., 2023a).

In the following section, we illustrate the effectiveness and scalability of the *NativQA* framework by showcasing steps to create the large-scale and multilingual Multi*NativQA* dataset.

3 Multi*NativQA* Dataset

Our Multi*NativQA* dataset encompasses 7 languages, ranging from high- to extremely low-resource on 7 different location/cities, covering 18 predetermined topics (see Table 1 for details). Multi*NativQA* captures linguistic diversity, by including several dialects for dialect-rich languages like Arabic.[2] We also added two linguistic variations of Bangla to reflect differences between speakers in Bangladesh and West Bengal, India. Furthermore, we opted to cover English queries from Dhaka and Doha, where English is commonly used as a second language.

3.1 Implementing *NativQA* Framework

Query Collection: For multilingual query collection, we started with various predetermined topics (see Table 2) derived from common concepts in everyday lives of users. Next, we asked the native speakers to write 10 to 50 queries per topic focusing on issues they encounter in their major cities

[2] In addition to Modern Standard Arabic (MSA), used formally and officially, we incorporated six Arabic dialects: Egyptian, Jordanian, Khaliji, Sudanese, Tunisian, and Yemeni, representing the diverse linguistic landscape of Doha, Qatar.

and then extended to urban areas. As there was no strict limit on the number of queries, some topics exceeded 50 queries. Then, we prompted GPT-4 to generate 10 similar queries based on each input query. The resultant number of seed queries, after de-duplication, for each language are reported in Table 1.

QA Collection: Consequently, we use the *QAC Module* to enrich queries and QA pairs for each language and its respective city. For each language, we ran our collection algorithm for 3-7 iterations (N_{iter}) based on the convergence rate. After the QAC, we collected ∼154K QA pairs across all languages as presented in Table 1 (column #QA).

QA Validation: The *QAV* is the final step of the *NativQA* framework. It includes two sub-steps: DRC and QAA. We implemented DRC across all target languages and cities. As for QAA, we only implement this sub-step for the evaluation (test) split of the Multi*NativQA* dataset. This is because the QAA is often time-consuming and costly but highly effective for designing evaluation datasets. However, due to resource constraints, we annotated the Arabic and Bangla datasets for QAA, and annotations for other languages are currently ongoing.

Query Topics
Animal, Business, Cloth, Education, Events, Food & Drinks, General, Geography, Immigration Related, Language, Literature, Names & Persons, Plants, Religion, Sports & Games, Tradition, Travel, Weather

Table 2: Selected topics used as seed to collect manual queries.

3.2 Annotation Guidelines

3.2.1 Domain Reliability

The objective for the domain reliability annotation task is to verify the credibility of the source domain, which can be used to judge the factuality and reliability of answers sourced from that domain. We adopt the following definition of the credibility of the domain/website: "A credible webpage is one whose information one can accept as the truth without needing to look elsewhere. If one can accept information on a page as true at face value, then the page is credible; if one needs to go elsewhere to check the validity of the information on the page, then it is less credible" (Schwarz and Morris, 2011).

Annotators were tasked to review each web domain to determine its credibility and assign one of the following four reliability labels:

- **Very reliable:** The information is accepted without additional verification.
- **Partially reliable:** The information may need further verification.
- **Not sure:** Unable to verify or judge the website for any reason.
- **Completely unreliable:** The website and the information appear unreliable.

In Section A.1 (in Appendix), we provide additional details of the instructions.

3.2.2 QA Annotation

This phase of the *NativQA* framework involves three types of annotations. Below, we discuss the guidelines for each type.

1. Question selection: The purpose of this task is to evaluate the quality of the questions. The annotators assessed whether the questions are factual or meet the criteria discussed below. We defined two types of questions inspired by the NQ dataset (Kwiatkowski et al., 2019).

- **Good question:** A good question is a fact-seeking question that can be answered with an entity or explanation.
- **Bad question:** A question is considered a bad question if it meets any of the following criteria:
  - Ambiguous or based on a false presupposition, making it incomprehensible.
  - Opinion-seeking, such as *"Can you give me your thoughts on. . . ?"*
  - Does not ask for factual information.

Based on whether a question is marked as good or bad, the annotator's subsequent tasks will vary. If a question is marked as good, the annotator will review the answer, its source page, and perform answer categorization tasks. Otherwise, further annotation is skipped, and the annotator proceeds to the next QA pair.

2. Answer categorization: An answer can be categorized into one of these categories: *(i)* correct answer, *(ii)* partially correct answer, and *(iii)* incorrect answer, and *(iv)* the answer can't be found in the source page. Complete definition for each category is provided in Section A.2.

3. Answer editing: The purpose of this step is to ensure that the answer accurately responds to the question and is correct, fluent, and informative. If the answer was incomplete, annotators are required to check the answer source page and extract content from it that can complete the answer (if such content is available in the page).

3.3 Annotation Task Setup

The annotation team consists of native speakers of the respective languages, who worked on the entire process starting from query collection to QA pair annotation. The annotators have diverse educational backgrounds, ranging from undergraduate students to those holding undergraduate or graduate degrees. The team was trained and monitored by an in-house expert annotator. To ensure quality, periodic checks of random annotation samples were conducted, and feedback was provided. Depending on the availability of the annotators for a language, we opted to go for one to three annotators for the domain reliability task. For the DRC task, three annotators were assigned to Arabic, Bangla, and English, while Assamese and Nepali each had one. When multiple annotators label a domain, the majority label is used as its final label. For other languages, domains were automatically matched with those already identified as reliable by annotators. For the QAA task, Arabic and Bangla, each QA pair was annotated by two annotators, while Assamese and Nepali each had one.

3.4 Annotation Platform

We utilized our in-house annotation platform for the annotation task. Separate annotation interfaces (as presented in Section B) were designed for each phase and each language. To facilitate the annotation process, the annotation interface included the annotation guidelines throughout the phases.

3.5 Annotation Agreement

To evaluate the reliability of manual annotations, we computed the Inter-Annotator Agreement (IAA) using a Fleiss' Kappa coefficient (κ) for the domain reliability task for Arabic, Bangla, and English. The Kappa (κ) values for these languages are 0.53, 0.66, and 0.37, respectively, which correspond to fair to substantial agreement according to Landis and Koch's scale (Landis and Koch, 1977). Note that we selected the final label where the majority agreed, meaning that we have above 66% agreement on the final label.

For the QA annotation task, we have two scenarios: *(i)* For languages with two annotators, we first directly select only the questions where both annotators agree. For the disagreed cases, another annotator revises them; ultimately, we select based on the agreement of at least two annotators. For the answer editing, 75.79% (Bangla) to 88.5% (Arabic) of the cases were agreed upon by both annotators, computed based on exact string matching; *(ii)* Languages with single annotator, we directly relied on their annotations.

Lang	Train	Dev	Test	Total
Arabic	5,284	747	1,517	7,548
Assamese	–	–	572	572
Bangla-BD	7,507	1,062	2,155	10,724
Bangla-IN	7,678	1,086	2,205	10,969
English-QA	8,964	1,268	2,574	12,806
English-BD	4,952	701	1,422	7,075
Hindi	9,604	1,358	2,758	13,720
Nepali	–	–	581	581
Turkish	5,619	795	1,613	8,027
Total	**49,608**	**7,017**	**15,397**	**72,022**

Table 3: Data split distribution for different languages.

3.6 Statistics and Analysis

In Figure 1, we report the initial collection of data distribution across languages, irrespective of the country they were collected from. English, Arabic, and Bangla are higher in proportion due to the fact that *(i)* English consists of data collected from Qatar and Bangladesh, *(ii)* Arabic consists of queries from different dialects, and *(iii)* Bangla consists of data from Bangladesh and India. As table 1 shows, our annotation process resulted in a decrease in QA set size by half (comparing initial QA set (column *#QA*) to final QA set (column *F.QA*)). We also faced a significant drop for Assamese and Nepali. This drop is due to the fact that the search engine returned QA pairs in non-native languages (in these cases, either Hindi or English) rather than the native language. As part of our process, we filtered out QA pairs that are not in the target language. We identify the native language using a language detection tool[3] and then manually revise them.

In Figure 10 and 11 (in Appendix), we report topic wise distribution for all languages and regions. It appears that we have a very good coverage of topics for all languages.

4 Experimental Setup

To establish baselines over Multi*NativQA*, we benchmark LLMs performance in the QA task over

[3] http://fasttext.cc/docs/en/language-identification.html

Models	BLEU	Rou.	MET.	BLEU	Rou.	MET.	BLEU	Rou.	MET.	BLEU	Rou.	MET.	BLEU	Rou.	MET.
	Arabic			**Bangla-IN**			**English-BD**			**Hindi**			**Turkish**		
GPT-4o	**0.315**	**0.052**	**0.107**	0.248	**0.019**	**0.108**	**0.325**	0.276	0.231	**0.337**	**0.056**	**0.168**	**0.305**	**0.247**	**0.131**
GPT-4	0.275	0.039	0.090	0.225	0.021	0.093	**0.325**	**0.278**	**0.234**	0.296	0.048	0.148	0.288	0.228	0.119
Gemini	0.239	0.037	0.102	**0.291**	0.015	0.014	0.202	0.230	0.217	0.288	0.044	0.167	0.259	0.226	0.124
LLama-3	0.143	0.026	0.054	0.051	0.004	0.047	0.285	0.215	0.183	0.112	0.027	0.080	0.098	0.145	0.067
Mistral	0.190	0.016	0.050	0.148	0.005	0.049	0.012	0.016	0.017	0.023	0.005	0.015	0.054	0.046	0.031
	Assamese			**Bangla-BD**			**English-QA**			**Nepali**					
GPT-4o	**0.110**	0.028	0.110	0.231	**0.012**	**0.103**	**0.333**	**0.268**	0.216	**0.300**	0.005	**0.114**			
GPT-4	0.086	0.019	0.085	0.210	0.011	0.085	0.331	**0.268**	**0.217**	0.280	0.004	0.090			
Gemini	0.107	**0.500**	**0.221**	**0.271**	0.009	0.095	0.236	0.241	0.212	0.114	**0.071**	0.111			
LLama-3	0.016	0.004	0.040	0.053	0.002	0.045	0.287	0.219	0.183	0.058	0.000	0.048			
Mistral	0.099	0.004	0.032	0.143	0.004	0.049	0.039	0.039	0.032	0.006	0.002	0.009			

Table 4: Reported results for different languages with different LLMs. MET.: METEOR, Rou.: Rouge1. **Bold** results are best per column per language.

it. Our experiments setup is described next.

Data Splits. The dataset for each language is split into training, development, and test sets using stratified sampling, considering topics as labels, with proportions of 70%, 10%, and 20%, respectively. Due to the small size of the Nepali and Assamese sets, we use their full datasets as a testing sets. Details of the final splits are in Table 3.

Models. We experiment with both open and close LLMs. For the close models we use GPT-4o, GPT-4 (version 0314) (Achiam et al., 2023), and Gemini[4]. For open models, we opt for llama-3-8b-instruct[5], and mistral-7b-instruct[6]. We use zero-shot learning as our setup with all models. For reproducibility, we set the temperature to zero, and designed the prompts using concise instructions.

Evaluation Metrics. We measure the performance the models using standard metrics commonly used for QA evaluation, such as BLEU, ROUGE and Meteor.

5 Results

Open *vs* Close LLMs Table 4 shows the complete results for each LLM and language. We observe that closed models, especially GPT-4o, clearly outperform all other models across majority of languages[7] with an average BLEU score of 0.278. This performance is then followed by GPT-4 and Gemini, with BLEU scores of 0.258 and 0.223, respectively. Results also show that in terms of open models, considering BLEU scores, Mistral is outperforming LLama3 in majority of the languages such as Arabic, Assamese, and Bangla (BD, IN).

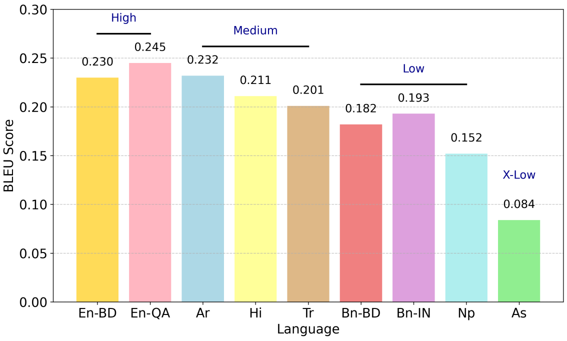


Figure 5: Average BLEU scores by language. X-Low: Extremely low.

High- *vs* Low-resource Languages We also look at the the average performance of models per language (Reporting average BLEU scores in Figure 5). The average performance over English surpasses that over other languages, which is expected given that English is the highest resource language among those we consider. Medium-resource languages such as Arabic, Turkish, and Hindi rank just below English in terms of performance. Assamese, categorized as an extremely low-resource language, exhibits very poor performance. Overall, the representation and/or richness of digital content for a language is reflected in the performance of the models.

6 Related Work

LLMs have consistently showcased impressive capabilities spanning diverse disciplines and tasks and there have been efforts to evaluate the perfor-

[4] gemini-1.5-flash-preview-0514

[5] https://ai.meta.com/blog/meta-llama-3/

[6] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1

[7] Except Bn-BD and BN-IN

mance of LLMs on standard NLP tasks (Achiam et al., 2023; Touvron et al., 2023b; Bubeck et al., 2023; Bang et al., 2023; Ahuja et al., 2023; Hendy et al., 2023). In a variety of domains, including finance (Wu et al., 2023), law (Liu et al., 2023), medicine (Wang et al., 2023a; Singhal et al., 2023), programming (Li et al., 2022) and intellectual property (Ni et al., 2024). NLP tasks has also been benchmarked in many studies. For example, Lai et al. (2023a) evaluated ChatGPT by considering seven different NLP tasks and covering 37 diverse languages with high, medium, low, and extremely low resource settings. While there have been several efforts to develop resources and benchmark LLMs with those resources, most of the prior works are limited to English. Furthermore, regarding the evaluation for other languages, translated forms are commonly used. With our effort we aimed to address this gap by proposing a framework (*NativQA*) and a dataset.

Existing QA Datasets There are many QA datasets in different languages. Below, we discuss the most widely used datasets. Kwiatkowski et al. (2019) and Yang et al. (2018) proposed two extractive QA datasets including Natural Questions(NQ), both containing long-form answers for questions that include large-scale question-answer pairs. The generated long answer's faithfulness is estimated by measuring the ratio of the golden short answer span contained in it. Joshi et al. (2017) developed TriviaQA dataset, which consists of 650K question-answer-evidence triples. These triples are created by merging 95K question-answer pairs. Rajpurkar et al. (2016) developed SquAD, which is a collection of 100K crowdsourced questions and answers paired with shortened Wikipedia articles. To develop the dataset, paragraphs/passages were given to annotators, and they were asked to write QA pairs based on the passage. HelpSteer (Wang et al., 2023b) is another QA dataset, which comprises a 37K sample dataset with multiple attributes of helpfulness preference that covers verbosity, accuracy, coherence, and complexity in addition to overall helpfulness.

Evaluations of LLMs for QA For benchmarking there are many notable datasets covering world knowledge (Hendrycks et al., 2020), commonsense reasoning (Zellers et al., 2019), reading comprehension (Bandarkar et al., 2023), factuality (Lin et al., 2022), and others. These datasets are usually transformed into multiple-choice questions. Additionally, standard QA datasets have also been used for LLM evaluation (Hu et al., 2020). Kamalloo et al. (2023) performed the analysis of different open-domain QA models, including LLMs by manually judging answers on a benchmark dataset of NQ-open (Lee et al., 2019), and reported a systematic study of lexical matching. Their investigation shows that LLMs attain state-of-the-art performance but fail in lexical matching when candidate answers become longer.

Dataset	# of Lang	Lang	Domain	Size
NQ (Kwiatkowski et al., 2019)	1	En	Wiki	323K
HotpotQA (Yang et al., 2018)	1	En	Wiki	113K
TriviaQA (Joshi et al., 2017)	1	En	Wiki, Web	650K
SquAD (Rajpurkar et al., 2016)	1	En	Wiki	100K
HelpSteer (Wang et al., 2023b)	1	En	Helpfulness	37K
BanglaRQA (Ekram et al., 2022)	1	Bn	Wiki	3k
Multi*NativQA* dataset	7	Ar, As, Bn, En, Hi, Np, Tr	Open	72K

Table 5: Existing most notable QA datasets in compare to ours (Multi*NativQA*).

In Table 5, we report the most notable existing QA datasets along with ours. Compared to existing datasets, Multi*NativQA* dataset is novel in terms of the number of languages, wide coverage of topics, and being native to the region and culturally aligned.

7 Conclusions

In this study, we propose the *NativQA* framework, which enables constructing culturally and regionally-aligned natural QA datasets. Resulting datasets can aid in training/fine-tuning and evaluating LLMs over native and culturally-aligned real users information needs and tasks. The proposed framework is scalable and reduces human involvement in a dataset construction by automating several processes. We show the efficacy of the *NativQA* framework, by designing and developing a multilingual native QA dataset, Multi*NativQA*. We further enrich our study by benchmarking QA performance of five open and closed LLMs over Multi*NativQA*. Our results demonstrate that the latest closed model, GPT-4o, shows superior performance in general across languages. Our study is an ongoing effort; therefore, we aim to extend the framework to include more languages and implement more measures to improve the quality of both the framework and the dataset. In future, the dataset will be used to tune LLMs to improve cultural and regional alignment.

8 Limitations

While the proposed framework enables the development of datasets with cultural and native information, it currently has several limitations. Firstly, the *NativQA* framework still relies on human-in-the-loop processes, from seed query creation to manual revision of QA pairs. This dependency limits large-scale data collection. Although we consider the human-in-the-loop setting a limitation, we also note that ensuring a high-quality dataset without it would be challenging. Secondly, the semi-supervised approach to dataset development is a reasonable starting point; however, full supervision would ensure higher quality. Thirdly, in our current study, we relied on one search engine. This can be extended to include other search engines and use a mixture of engines to enrich QA pair collection. Fourth, due to resource limitations, including the availability of language-specific annotators, we have successfully annotated QA pairs for the test sets in Arabic and Bangla. Annotation for other languages including development and training sets is a part of our ongoing efforts. Finally, our study is currently limited to benchmarking various open and closed models. Future research will focus on fine-tuning and training new models.

Ethics and Broader Impact

The proposed *NativQA* framework does not involve collecting any personally identifiable information. Additionally, the proposed dataset does not include any information that can offend or harm any individual, entity, organization, or society. Therefore, we do not foresee any issues that may lead to potential risks. Human annotators were paid through external companies at standard payment rates applicable to their region. Information about human annotators is not part of the dataset, and their identities remain confidential. The proposed framework and dataset will be released publicly for non-commercial research purposes. Therefore, we strongly believe that they will be beneficial for the research community.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675—-718, Indonesia. Association for Computational Linguistics.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. Technical report, Microsoft Research.

Syed Mohammed Sartaj Ekram, Adham Arik Rahman, Md. Sajid Altaf, Mohammed Saidul Islam, Mehrab Mustafy Rahman, Md Mezbaur Rahman, Md Azam Hossain, and Abu Raihan Mostofa Kamal. 2022. BanglaRQA: A benchmark dataset for under-resourced Bangla language reading comprehension-based question answering with diverse question-answer types. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2518–2532, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Andrew J. Flanagin and Miriam J. Metzger. 2007. The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. *New Media & Society*, 9(2):319–342.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan

Awadalla. 2023. How good are GPT models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

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 generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Ziche Liu, et al. 2023. Acegpt, localizing large language models in arabic. *arXiv preprint arXiv:2309.12053*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hiéu Mãn, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023a. 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.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023b. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv preprint arXiv:2307.16039*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. 2022. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Hongcheng Liu, Yusheng Liao, Yutong Meng, and Yuhao Wang. 2023. Lawgpt: Chinese legal dialogue language model.

Miriam J Metzger and Andrew J Flanagin. 2015. Psychological approaches to credibility assessment online. *The handbook of the psychology of communication technology*, pages 445–466.

Shiwen Ni, Minghuan Tan, Yuelin Bai, Fuqiang Niu, Min Yang, Bowen Zhang, Ruifeng Xu, Xiaojun Chen, Chengming Li, and Xiping Hu. 2024. MoZIP: A multilingual benchmark to evaluate large language models in intellectual property. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11658–11668, Torino, Italia. ELRA and ICCL.

OpenAI. 2023. GPT-4 technical report. Technical report, OpenAI.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Julia Schwarz and Meredith Morris. 2011. Augmenting web pages and search results to support credibility assessment. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1245–1254.

Ovidiu Selejan, Dafin F Muresanu, Livia Popa, I Muresanu-Oloeriu, Dan Iudean, Arica Buzoianu, and Soimita Suciu. 2016. Credibility judgments in web page design–a brief review. *Journal of medicine and life*, 9(2):115.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024. Kmmlu: Measuring massive multitask language understanding in korean. *arXiv preprint arXiv:2402.11548*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023a. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

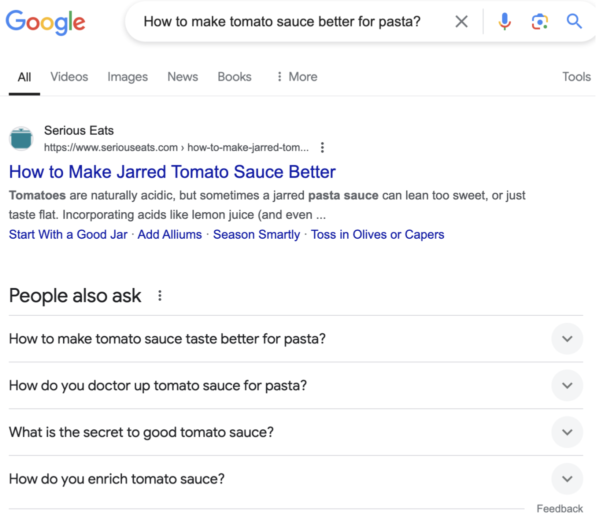
Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023a. Huatuo: Tuning llama model with chinese medical knowledge. *CoRR*, abs/2304.06975.

Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Polak Scowcroft, Neel Kant, Aidan Swope, et al. 2023b. Helpsteer: Multi-attribute helpfulness dataset for steerlm. *arXiv preprint arXiv:2311.09528*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David S. Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *CoRR*, abs/2303.17564.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.



How to make tomato sauce better for pasta?

Serious Eats
https://www.seriouseats.com › how-to-make-jarred-tom...

How to Make Jarred Tomato Sauce Better

Tomatoes are naturally acidic, but sometimes a jarred **pasta sauce** can lean too sweet, or just taste flat. Incorporating acids like lemon juice (and even ...

Start With a Good Jar · Add Alliums · Season Smartly · Toss in Olives or Capers

People also ask

How to make tomato sauce taste better for pasta?

How do you doctor up tomato sauce for pasta?

What is the secret to good tomato sauce?

How do you enrich tomato sauce?

Figure 6: An example of search interface showing search response with *"people also ask"* option.

Appendix

A Query on Search Engine

In Figure 6, we show an example of a query to a search engine, that demonstrates related queries under "People also ask", which we have also considered as queries in the several iterations of QA pair collection.

A.1 Domain Reliability

General Characteristics Below are the characteristics that we have considered as criteria for a domain to be more reliable:

Overall Design:

- The domain has a professional, polished, and attractive design. It has interactive features,

- is well organized, easy to navigate, loads fast, and has good response speed.
- There are no errors or broken links.
- It might have paid access to information.
- The domain name suffix is considered trustworthy (e.g., ".gov").
- Absence/limited advertising. If advertising is present, they are good quality ads for reputable and decent products and organizations.
- The domain might be sponsored by or shows links to reputable organizations.
- Presence of privacy and security policies section or page. Presence of an About page, contact info, and address.
- If videos, images, and graphics are used on the website, they are high-quality and professional.

Content Quality:

- Author/entity names, qualifications, credentials, and contact information are present, and they are relevant to the topic of the content.
- Author/entity is reputable.
- Contains date stamp
- Presents information that is current and up to date.
- Has citations, especially to scientific data or references, and shows links to external authorities.
- Content is relevant to the target topic and current events.
- Professional-quality, clear writing, good formatting of text.
- Content appears accurate, lacks bias, factually correct, plausibility, and uses appropriate objective language.
- Free of misspellings, grammar mistakes, etc.
- The information provided is at an appropriate level, not too generic or elementary.

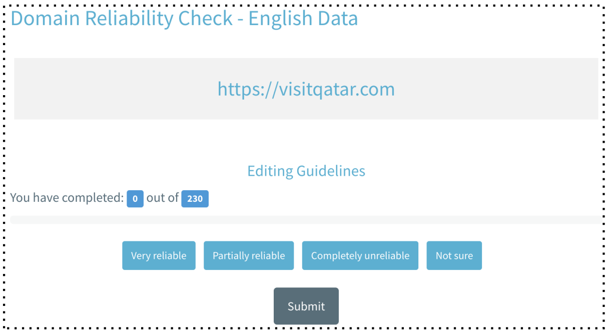
General Instructions: We also provided the following general instructions to guide the annotation process.

- Do not spend more than five minutes per given Web domain.
- Explore/observe/look at **ALL** elements in the domain's home page from top to bottom.
- Repeat points 1-2 on other pages from the same domain, and look at their content, structure, design, author, etc. *You are not required to read these pages in full, reading the first 1-2 paragraphs is enough.*
- During annotation, consider the annotation criteria mentioned in this guideline, and evaluate each source based on those aspects. A "reliable website" might not meet all those criteria. It is your job to measure the website's reliability guided by these elements.
- You should evaluate a source based on that source only and what it presents. You should not navigate or search for outside sources even if some are linked inside the given domain/-page.
- Please use "Not sure" very sparingly in rare cases when you are extremely unsure. It is preferable to always choose one of the other three labels.
- For social media websites (e.g., X, Facebook) choose: Very Reliable.
- For shopping websites, use the criteria listed in this guideline to decide. Some shopping websites are very reliable.
- For famous people's websites, use the criteria listed in this guideline to decide.
- Websites that are in any other language ONLY (for example, only in En when you are working on a Bangla queries), for such cases choose: Not Sure.

A.2 Definitions of Answer Categorization

Below we provide the definition of the categories that are defined for the answer categorization task.

- **Correct answer:** When the answer aligns with the information provided by the source. Note that the answer must be complete and address all parts of the question, but it does not need to match the source webpage verbatim. The answer can be a long, detailed response or a short snippet.
- **Partially correct answer:** When the answer does not address all parts of the question. In this case, the goal is to edit the answer using information from the source page. This involves directly copying text from the source webpage. Minimal editing may be needed to make the answer more comprehensive.
- **Incorrect answer:** When the answer text does not address the question. In this case, the goal is to edit the answer using information from the source page.
- **Cannot find answer:** When the answer is not available on the provided link/page.



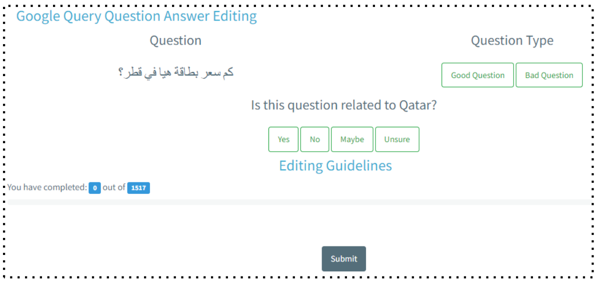
Domain Reliability Check - English Data

https://visitqatar.com

Editing Guidelines

You have completed: 0 out of 230

Figure 7: An example of the annotation interface for domain reliability checking.



Google Query Question Answer Editing

Question: كم سعر بطاقة هيا في قطر؟

Question Type

Is this question related to Qatar?

Editing Guidelines

You have completed: 0 out of 1517

Figure 8: Annotation interface for *Question selection*.

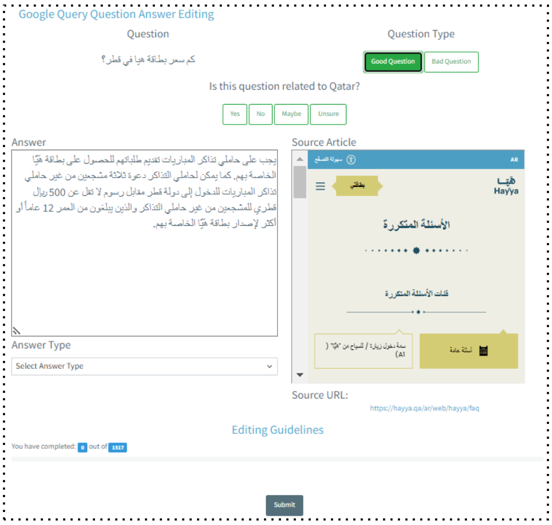
B Annotation Interface

In Figure 7, we present an example of domain reliability checking, which consists of a URL of the domain, annotation guidelines, and four different options associated with the four categories we defined for this annotation task. Annotators select one of these options and submit.

In Figure 8 and 9 we demonstrate the two steps of question selection and answer editing and categorization tasks, respectively. Depending on the type of question selected, the annotator will be able to choose whether to edit the answer or not.

C Dataset: Additional Details

In Figure 10 and 11 we present the topic-wise data distribution for different datasets associated with various languages. Starting with the Arabic dataset, the predominant topic is *names*, comprising 10.2% of the data, a trend that also holds true for Assamese (8.2%). For Bangla, whether from Bangladesh or India, the major topic is *general*, representing 8.8% and 10.0% respectively. In Bangladesh, *religion* (10.4%) is the major topic for English, whereas in Qatar, *general* dominates at 26.6%. For Nepali, the leading topic is *Business* (22.9%), for Hindi it is *Travel* (8.2%), and for Turkish, *names* is the primary topic at 8.6%.



Google Query Question Answer Editing

Question: كم سعر بطاقة هيا في قطر؟

Question Type: Good Question

Is this question related to Qatar?

Answer:
يجب على حاملي تذاكر المباريات تقديم طلباتهم للحصول على بطاقة هيّا
الخاصة بهم. كما يمكن لحاملي التذاكر دعوة ثلاثة مشجعين من غير حاملي
تذاكر المباريات للدخول إلى دولة قطر مقابل رسوم لا تقل عن 500 ريال
قطري للمشجعين من غير حاملي التذاكر والذين يبلغون من العمر 12 عاماً أو
أكثر لإصدار بطاقة هيّا الخاصة بهم.

Answer Type: Select Answer Type

Source Article

Source URL: https://hayya.qa/ar/web/hayya/faq

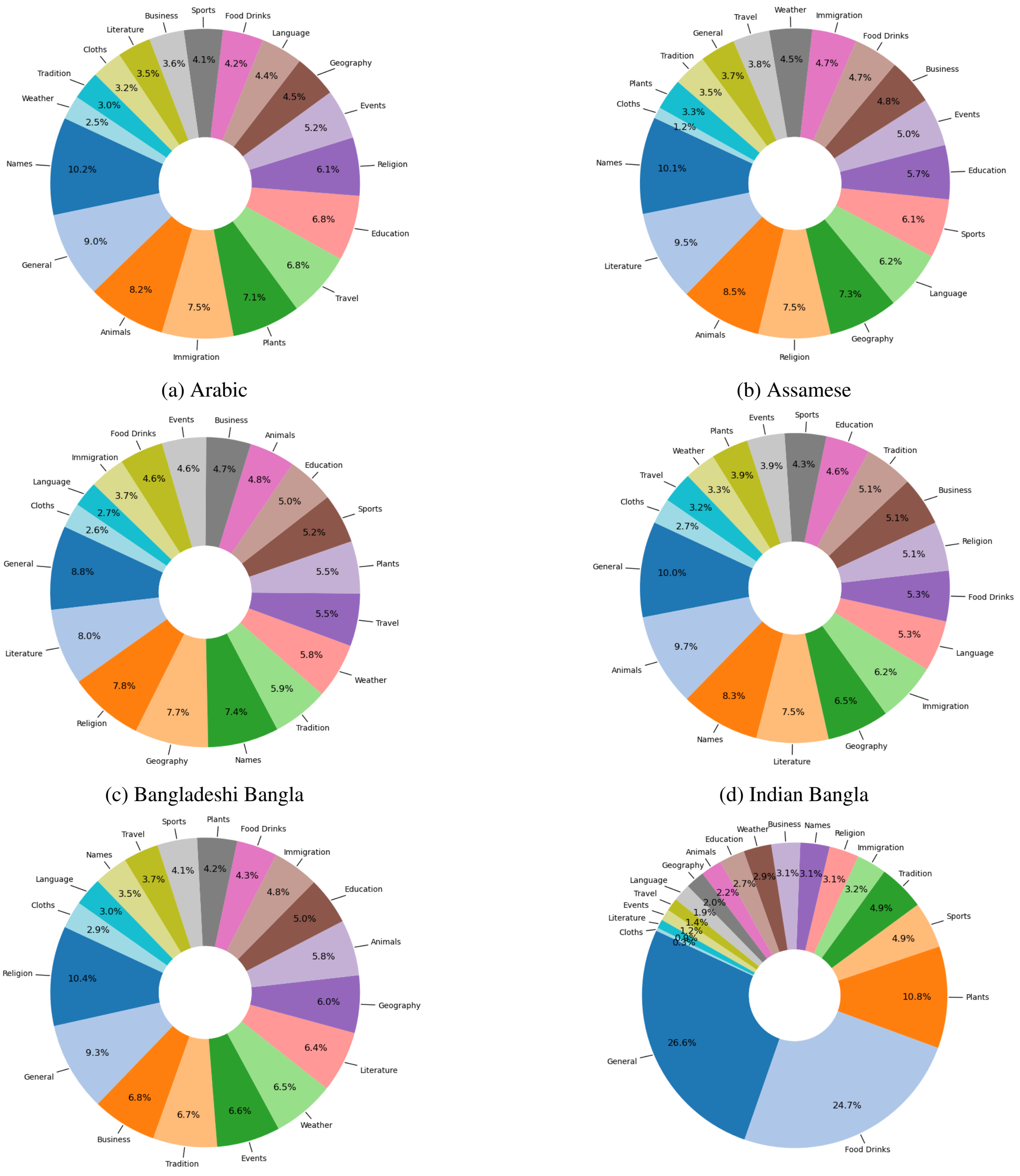
Editing Guidelines

You have completed: 0 out of 1517

Figure 9: Annotation interface for *Answer editing* and *answer categorization*.

D Data Release and License

The *NativQA* dataset will be publicly released under the Creative Commons Attribution Non Commercial Share Alike 4.0: `https://creativecommons.org/licenses/by-nc-sa/4.0/`.



(a) Arabic

(b) Assamese

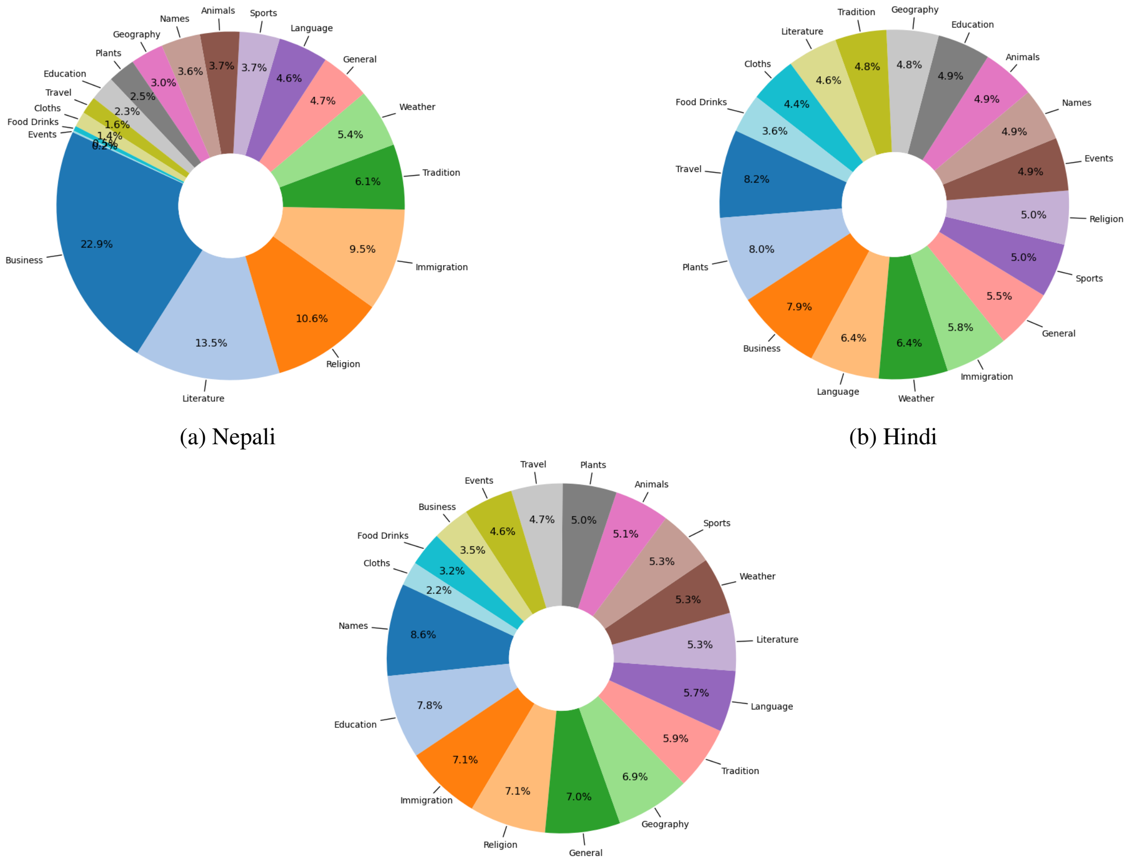
(c) Bangladeshi Bangla

(d) Indian Bangla

(e) English in Bangladesh

(f) English in Qatar

Figure 10: Topic wise distribution in different languages such as *Arabic*, *Assamese*, *Bangladeshi Bangla*, *Indian Bangla*, *English in Bangladesh*, and *English in Qatar*.



(a) Nepali

(b) Hindi

(c) Turkish

Figure 11: Topic wise distribution in different languages such as *Nepali*, *Hindi* and *Turkish*.