TyDi QA–WANA: A Benchmark for Information-Seeking Question Answering in Languages of West Asia and North Africa

Parker Riley
Google

Siamak Shakeri
Google

Waleed Ammar[*]
Holistic Intelligence for Global Good

Jonathan H. Clark
Google

Abstract

We present TyDi QA–WANA, a question-answering dataset consisting of 28K examples divided among 10 language varieties of western Asia and northern Africa. The data collection process was designed to elicit *information-seeking* questions, where the asker is genuinely curious to know the answer. Each question in paired with an entire article that may or may not contain the answer; the relatively large size of the articles results in a task suitable for evaluating models' abilities to utilize large text contexts in answering questions. Furthermore, the data was collected directly in each language variety, without the use of translation, in order to avoid issues of cultural relevance. We present performance of two baseline models, and release our code and data to facilitate further improvement by the research community.

1 Introduction

Many users worldwide regularly use technology to help them answer *information-seeking* questions, where the user does not know the answer but wants to. When developing systems to answer these questions, improving quality requires being able to reliably measure it using trustworthy and challenging evaluations. For measuring performance on answering English questions, multiple evaluation datasets are available. Some datasets are also available in non-English language varieties, but they cover only a tiny portion of the language varieties of the world.

Modern LLMs have shown promising performance in QA tasks, including non-English ones (Gemini Team, 2024b). However, performance is known to be worse for low-resource language varieties for which these models have seen comparatively little pretraining data. This highlights the need for training and evaluation data in these languages.

Additionally, the advent of long-context LLMs, which can accept more than 1 million input tokens (Reid et al., 2024), poses a challenge in evaluation because most information-seeking datasets are not designed to test a model's ability to utilize such a large context window.

We seek to address these issues by creating and releasing[1] a dataset of long-context information-seeking questions in under-represented non-English language varieties of West Asia and North Africa, in the style of TyDi QA (Clark et al., 2020), which we call TyDi QA–WANA. We also present baseline results that illustrate that modern LLMs are capable of answering questions by including *an entire Wikipedia article* in the input, something made feasible by recent advances in long-context modeling.

The rest of this work is organized as follows. First, we define the long-context information-seeking question answering task (§2). Next, we describe how we created the data (§3). We then outline the gap in prior work that this work addresses (§4) and catalog the language varieties selected for inclusion in our dataset (§5). To quantify these contributions, we present basic statistics of our dataset (§6), describe an evaluation procedure (§7), and present results on baseline systems (§9).

2 Task Definition

In this work we adopt a single task from Clark et al. (2020): the **Minimal Answer Span Task** (abbreviated as MinSpan). In this task, the input is the full text of an article paired with a question, and the desired output is one of three types:

1. The start and end byte indices of the minimal

[*] Work done at Google.

[1] Our code and data are available at https://github.com/google-research-datasets/tydiqa-wana.

span that completely answers the question, if such a span exists within the article.

2. YES or NO if the question requires a yes/no answer and such a conclusion can be drawn from the article.

3. NULL if it is not possible to answer the question with one of the above two types.

3 Data Collection Procedure

We similarly adopt the data collection procedure of Clark et al. (2020), which consists of question elicitation, article retrieval, and answer labeling.

Question Elicitation Our goal is to elicit *information-seeking* questions, where the annotator asking the question is actually interested in the answer. To this end, annotators are shown short prompts consisting of the first 150 characters from Wikipedia articles in their language variety and asked to write questions that are not directly answered by the prompt and that the annotator is curious about. The intent behind the prompts is to inspire the raters to inquire about various topics; note however that we do not require that the generated questions actually pertain to the provided prompts. Clark et al. (2020) provide a detailed justification for why it is important to elicit questions not directly answered by the prompts; briefly, this avoids questions that can be answered by simple techniques, resulting in a more challenging task.

Article Retrieval For each elicited question, we leverage a Google search restricted to the Wikipedia domain for its respective language variety[2] and retrieve the first result if any are found. If not found, the question is discarded and not included in our dataset. Note that we process the retrieved text to remove tables, long lists, and infoboxes, so that our resulting dataset is focused on natural text.

Answer Labeling Each found article is paired with its corresponding question and presented to 1 or 3 human annotators: 1 for our train split, 3 for our development and test splits. After first confirming that the question follows the guidelines discussed above, the annotator selects a paragraph in the article that contains the answer, or indicates that there is no such paragraph. The annotator then selects a **minimal answer**: a span of characters in the paragraph that is as short as possible while still comprising a satisfactory answer to the question. For example, for the question "Who was the first President of the United States?", annotators should select "George Washington" as the minimal answer. While many questions have a minimal answer of just a few words, others (such as "What is an atom?") require most of a sentence to answer.

We included multiple quality control and training steps in our data collection pipeline. For each rater, we first elicited a small number of English questions and provided feedback on any that did not meet our guidelines. Then we elicited a small number of in-language questions and had them reviewed by trusted in-house native speakers to verify that each rater is a native speaker of their respective language variety. In the answer labeling phase, we first had our raters review the task instructions and take a multiple-choice quiz covering how to annotate 21 example English question/article pairs, alternating between taking the quiz and reviewing the instructions until they achieved a score greater than 90%. After passing, raters performed answer labeling on a small number of English examples which we manually reviewed and provided feedback on; raters completed this stage repeatedly (with different examples each time) until we were satisfied with their task understanding. Our raters are paid fair market rates for their time, including training time.

4 Related Work

Many existing extractive question answering datasets evaluate *reading comprehension*, including SQuAD 2.0 (Rajpurkar et al., 2018), XQuAD (Artetxe et al., 2019), MLQA (Lewis et al., 2019), M2QA (Engländer et al., 2024), and LAReQA (Roy et al., 2020). The common feature of these datasets is that each question was written after reading the paired passage. We contrast this with *information-seeking* QA datasets, as in this work, where the question was written independently of the passage via a process designed to elicit questions that the question writer is curious about; examples include Natural Questions (Kwiatkowski et al., 2019) and its multilingual counterpart MKQA (Longpre et al., 2021), as well as TyDi QA (Clark et al., 2020).

Our dataset includes questions written in Arabic regional varieties with passages in Mod-

[2] For all Arabic varieties, only the questions are in that variety, while the articles are written in Modern Standard Arabic (see Section 5).

ern Standard Arabic. This is related to cross-lingual QA datasets such as XOR TyDi QA (Asai et al., 2020), XTREME-UP (Ruder et al., 2023), XQuAD (Artetxe et al., 2019), MLQA (Lewis et al., 2019), and LAReQA (Roy et al., 2020), though none of those evaluate cross-variety QA. Conversely, existing datasets examining Arabic dialects such as MSDA (Boujou et al., 2021), IADD (Zahir, 2022), and MADAR (Bouamor et al., 2019) are not QA datasets.

Our dataset includes Tajik and Azerbaijani, two low-resource language varieties. Examples of existing work in these varieties include Dovudov et al. (2012), Hecking and Sarmina-Baneviciene (2010), Merchant and Tang (2024), and Isbarov et al. (2024). None of these provided QA datasets for these varieties.

Our dataset emphasizes long-context modeling by requiring models to search for an answer within an entire article (see Table 1 for statistics on average article length). We expect this to be useful to evaluate the capabilities of models from recent years designed to process text of this length (Beltagy et al., 2020; Zaheer et al., 2020; Ainslie et al., 2020; Liu et al., 2024; Reid et al., 2024).

One motivation for our decision to elicit questions in-language instead of creating a parallel QA dataset by translating a fixed set of questions is to ensure that the resulting questions are culturally relevant within each language. This motivation is shared by CaLMQA (Arora et al., 2024), a QA dataset designed to be culturally relevant.

Our primary focus in this work is on developing an extractive QA dataset that: (1) contains information-seeking questions; (2) includes a set of languages that balances diversity (to assess coverage) and relatedness (to facilitate transfer learning); and (3) requires long-context capabilities. To our knowledge, ours is the first dataset to cover all of these criteria.

5 Selected Language Varieties

This section provides a brief description of each language variety included in our dataset, grouped by language family (and branches in some cases). IETF BCP 47 language tags are provided for all varieties.

The 10 language varieties in our dataset cover 3 different language families. Including multiple language families helps to assess models' language coverage, while including multiple related languages from within the same family allows practitioners using our dataset to evaluate transfer learning.

5.1 Afro-Asiatic Language Family

6 varieties from this language family are included.

5.1.1 Arabic

The *questions* in our dataset include 4 varieties of Arabic, while the *articles* are all from Arabic Wikipedia which is written in a fifth variety (Modern Standard Arabic). All varieties use the Arabic alphabet.

Modern Standard Arabic (*ar*) is primarily a literary variety of Arabic, and not generally spoken as a first language.

Algerian Arabic (*arq*) is a variety of Arabic spoken primarily in Algeria.

Egyptian Arabic (*arz*) is a variety of Arabic spoken primarily in Egypt. It is the most widely-spoken Arabic variety.

Iraqi Arabic (*acm*) is a variety of Arabic spoken primarily in Iraq, Syria, Turkey, Iran, and Kuwait.

Jordanian Arabic (*apc-JO*) is a variety of Arabic spoken primarily in Jordan.

5.1.2 Northwest Semitic

Hebrew (*he*) is primarily spoken in Israel and is written in the Hebrew script.

5.2 Indo-European Language Family

Armenian (*hy*) is spoken primarily in Armenia and written in the Armenian alphabet. It is the only member of an independent branch of the Indo-European language family.

5.2.1 Persian

We include two mutually-intelligible Persian language varieties.

Farsi (*fa*), also known as Iranian Persian or just Persian, is spoken primarily in Iran and written in the Persian alphabet.

Tajik (*tg*), also known as Tajiki Persian or just Tajiki, is spoken primarily in Tajikistan and written in the Cyrillic alphabet.

5.3 Turkic Language Family

Our dataset includes two mutually-intelligible language varieties from the Turkic family.

Azerbaijani (*az*) is a variety spoken primarily in Azerbaijan. In Azerbaijan it is written in Latin script, though elsewhere it is sometimes written in Arabic, Cyrillic, or Georgian scripts. In our dataset, questions and articles are in Latin script.

Turkish (*tr*) is a variety spoken primarily in Türkiye (also known as Turkey) and written in Latin script.

6 Dataset Statistics and Examples

Our dataset contains a total of 28,197 questions, including 16,200 examples in the training split (with 1 annotation each), 5,995 dev examples (3 annotations each), and 6,002 test examples (3 annotations each). Table 1 lists basic statistics about our dataset, including the number of examples by language and split. It also lists the percentage of examples where a majority of annotations were NULL, meaning that the annotator did not find an answer in the article (see Section 7 for further discussion on treatment of NULL annotations). In many language varieties, the proportion of NULL examples is quite high. We believe that this is primarily due to these varieties' Wikipedias being comparatively small: the smaller number of articles makes it more likely to match a question with an irrelevant article, and the shorter average length of articles makes it less likely for the answer to be present.

Some examples from our dataset are shown in Figure 1. In the first Turkish example, two annotators agreed exactly on the answer while a third did not find an answer (NULL). In this case, the NULL annotation will be discarded in score calculation (see Section 7.1). In the second Turkish example, all 3 annotators agreed that the yes-no question's answer was YES, which is supported by the article (a relevant excerpt has been selected for the figure). In the first Azerbaijani example, there is disagreement among the annotators about whether to include the day (July 13) in the example alongside the year (1318). This illustrates that training raters to annotate in exactly the same way is difficult. In the second Azerbaijani example, the article does not contain the answer, but one annotator selected a plausible yet incorrect answer. This is an example of our use of 3 annotators for the evaluation set, along with the NULL consensus procedure, mitigating noisy annotations.

7 Evaluation

The primary metric for our dataset is average per-example F1, though we also report exact match (EM) scores. Both F1 and EM for each example are calculated as the maximum score against each of the three available annotations for that example, with one large exception related to the distinction between NULL and non-NULL answers. Details on this distinction, as well as the specifics of score calculation, are presented in this section.

7.1 NULL Consensus

For many questions in our dataset, some or all annotators were unable to find an answer in the presented article. This is especially true for some language varieties where the associated Wikipedia is relatively small. As argued in Clark et al. (2020), model performance may be artificially inflated if a model could get credit for predicting NULL when **any** annotation is NULL. Thus, we first establish a "NULL consensus" for each example in our evaluation splits: if fewer than two annotations are NULL, then any NULL annotations are discarded. Conversely, if at least two are NULL, then any non-NULL annotations are discarded. This essentially means that we take a majority vote to determine whether there are any valid answers to the question, which is intended to limit noise in the final labels. For the training split, where only one annotation is present, the consensus is NULL if and only if the provided annotation is NULL.

7.2 F1 Score Calculation

For each example, if the model produces YES, NO, or NULL, then the F1 score for that example is 1.0 if any non-discarded annotation (see Section 7.1) matched that answer, and 0.0 otherwise. If the model produces valid start and end byte indices (referring to a candidate minimal span answer within the article), the indices in that range are treated as a set and compared to the set of indices selected in each annotation that contains a minimal answer span: the F1 score against a single annotation is the harmonic mean of precision and recall over byte indices, and the score for the example is the maximum score over available annotations (0.0 if no annotations contained a minimal span answer), to account for annotator variability in selecting the minimal span. The per-example F1 scores are then averaged to produce the model's final score.

7.3 Exact Match Calculation

The EM score is the proportion of examples where the model's output exactly matched an available

Turkish

Question: *Kaç tür asit vardır?* (How many types of acids are there?)
Article: [...] *Asitler başlıca iki grupta toplanabilirler:* [...] (Acids can be divided into two main groups:)
Answer Annotations:

1. *iki* (two)
2. *iki* (two)
3. NULL

Turkish

Question: *rRNA katlanma yapar mı?* (Does rRNA fold?)
Article: [...] *Ribozomdaki proteinler rRNAńın belli kısımlarını tanıyıp oralara bağlanır, ardından bu proteinler arasındaki etkileşimler rRNA'nın daha da katlanmasına neden olur ve sonunda ribozom meydana gelir.* [...] (Proteins in the ribosome recognize and bind to certain parts of the rRNA, then the interactions between these proteins cause the rRNA to fold further, eventually forming the ribosome.)
Answer Annotations:

1. YES
2. YES
3. YES

Azerbaijani

Question: *Fəzullah Rəşidəddin nə zaman vəfat edib?* (When did Fazullah Rashiduddin die?)
Article: [...] *O, Elxani hökmdarı Olcaytunu zəhərləməkdə mühakimə olunaraq 13 iyul 1318-ci ildə, 70 yaşında edam edilmişdir.* [...] (He was tried for poisoning the Elkhanid ruler Oljait and executed on July 13, 1318, at the age of 70.)
Answer Annotations:

1. *1318* (1318)
2. *1318* (1318)
3. *13 iyul 1318* (July 13, 1318)

Azerbaijani

Question: *Ukraynada ilk konsitusiya nə vaxt qəbul olunub?* (When was the first constitution adopted in Ukraine?)
Article: [...] *28 iyun — Ukraynanın Konstitusiya günü.* [...] (June 28 — Constitution Day of Ukraine.)
Answer Annotations:

1. NULL
2. NULL
3. *28 iyun* (June 28)

Figure 1: Examples from our development set. The articles have been trimmed to the relevant portion for clarity. Minimal span answers are represented as byte ranges in our dataset, but in this figure, the text indicated by those ranges is shown. Note that the English glosses are not included in the dataset.

Language Variety	Train Examples (1 Rating)	Dev Examples (3 Ratings)	Test Examples (3 Ratings)	Avg. Question Tokens	Avg. Article Bytes	Avg. Answer Bytes	% With NULL Consensus
Algerian Arabic	1306	649	647	5.8	92K	145	69.6%
Egyptian Arabic	1519	754	755	6.3	131K	60	49.2%
Jordanian Arabic	1386	688	690	5.3	114K	84	50.7%
Iraqi Arabic	1320	658	657	6.6	91K	42	66.2%
Armenian	1673	823	824	7.0	119K	39	76.7%
Azerbaijani	1469	724	724	6.0	48K	18	68.1%
Hebrew	1517	760	759	6.4	75K	43	48.8%
Farsi	3275	105	104	8.1	73K	46	47.9%
Tajik	1114	86	88	5.9	37K	36	82.7%
Turkish	1624	749	754	5.2	36K	30	51.4%

Table 1: Basic data statistics. “Avg. Question Tokens” is the average number of whitespace-delimited tokens in questions. “Avg. Answer Bytes” is the average number of bytes of all minimal answers where the consensus is non-NULL. “% With NULL Consensus” is the percentage of examples where the consensus is NULL. These values are calculated over all examples from all splits.

annotation. It is calculated in exactly the same way as F1 except when comparing minimal span answers between the model and annotations: instead of assigning partial credit, the score is 1.0 if the byte ranges exactly match, and 0.0 otherwise. Thus, the EM score is never higher than the F1 score.

8 Baseline Systems

To provide baseline results as a point of comparison for future work on our dataset, we present results from two publicly-available models on our task: Gemini 1.5 Pro (Gemini Team, 2024a) and Gemini 2.0 Flash[3]. Both are large language models trained on multilingual text.

The input consists of three components: 1) The preamble, 2) one exemplar of each question type for in-context learning, and 3) the input question. For each language, a set of exemplars was collected and used for all samples within that language.

The following is the preamble used:

> *Consider the following question and passage, which may contain an answer. Find the answer, if it exists. If there is not a good and complete answer to the question, respond with "no answer". If the question is asking for a yes or no answer, return "yes" or "no" if there is evidence for this in the passage. Answers will usually be just a few words, but could be more in some cases.*

Each exemplar is formatted as follows:

> Article:<*article text*>
> Q:<*question text*>
> A:<*answer text*>

For yes/no answers, the answer text is “yes” or “no”, and for NULL answers, the answer text is “no answer”; evaluation is case-insensitive for these answer types. For minimal span answers, the task definition requires a byte range but off-the-shelf LLMs cannot reliably produce them, so instead the model produces the answer text directly and, as a heuristic, we locate the first occurrence of that text within the article to recover the byte range.

Different exemplars are concatenated by a newline. In order to reduce the total input length to the model, we selected exemplars from the shorter articles. We use greedy sampling with the maximum generation length of 1024.

[3] https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/

Gemini Model →	1.5 Pro		2.0 Flash	
Language ↓	EM	F1	EM	F1
Algerian Arabic	46.5	48.2	47.3	48.9
Egyptian Arabic	54.1	57.9	54.9	58.6
Iraqi Arabic	55.3	56.9	56.2	58.0
Jordanian Arabic	46.9	54.6	48.8	56.4
Armenian	74.0	74.8	72.7	73.6
Azerbaijani	71.8	74.4	71.5	74.3
Hebrew	50.0	52.8	50.5	52.7
Farsi	38.1	41.9	40.0	48.3
Tajik	73.3	74.2	66.3	67.1
Turkish	55.7	58.0	56.3	58.4

Table 2: Exact Match and F1 scores (both out of 100) of our baseline systems on all 10 language varieties on our development set.

Gemini Model →	1.5 Pro		2.0 Flash	
Language	EM	F1	EM	F1
Algerian Arabic	45.0	46.9	45.7	47.7
Egyptian Arabic	53.0	58.4	53.1	58.0
Iraqi Arabic	60.6	62.9	61.5	64.3
Jordanian Arabic	46.8	56.3	48.1	57.6
Armenian	71.8	72.7	71.1	72.2
Azerbaijani	74.9	76.2	73.6	75.1
Hebrew	53.4	57.3	52.6	56.3
Farsi	34.6	38.7	45.2	50.4
Tajik	68.2	68.2	56.8	57.6
Turkish	53.3	55.1	53.8	55.3

Table 3: Exact Match and F1 scores (both out of 100) of our baseline system on all 10 language varieties on our test set.

8.1 No-Answer Critic

While we direct the baseline models to output “no answer” when it cannot find an answer to the question, we observed that they sometimes produced responses indicating that no answer was found but in the wrong format, such as “I can’t find an answer to this question” (either in English or in the question’s language variety). Because we are not primarily interested in measuring the models’ adherence to this particular output format, we use a second inference pass with the model to analyze its own output to determine whether the answer is a misformatted version of “no answer”. We only use this no-answer critic when the model’s output is not a valid substring of the article, and also none of “yes”, “no”, or “no answer” (case-insensitive). We observed this critic to be very effective at detecting NULL cases that did not follow the requested output format. See Appendix A.1 for the prompt used for the no-answer critic.

9 Results and Discussion

Tables 2 and 3 report our baseline’s performance on the development and test splits, respectively. Neither baseline clearly outperforms the other overall. For both models, scores vary widely between language varieties. We note that this is not sufficient to conclude that a model is better at answering questions in e.g. Armenian than in Farsi: the questions are different, and specifically the proportion of NULL-consensus answers in Armenian is higher than in Farsi. We therefore recommend that practitioners using our dataset focus on comparing the scores of different models within the same language varieties to measure per-variety improvement, as opposed to comparing scores across varieties for the same model.

Gemini Model →	1.5 Pro		2.0 Flash	
NULL consensus →	✓	✗	✓	✗
Language ↓				
Algerian Arabic	57.0	16.9	57.8	17.2
Egyptian Arabic	63.2	53.1	64.3	53.3
Iraqi Arabic	60.3	48.8	61.6	49.4
Jordanian Arabic	58.7	50.2	61.5	51.0
Armenian	85.0	39.5	83.7	38.7
Azerbaijani	82.7	58.2	81.7	59.8
Hebrew	60.1	45.3	61.9	43.2
Farsi	20.0	45.5	6.7	55.2
Tajik	75.7	65.4	67.6	64.0
Turkish	61.6	54.2	61.9	54.7

Table 4: Baseline F1 results on our development set, divided into examples with a NULL (✓) vs. non-NULL (✗) consensus.

By comparing EM and F1 scores, we can estimate the frequency with which the baseline models predict an answer span that overlaps with, but does not exactly match, a gold answer span. The fact that EM and F1 scores are relatively close to each other in Tables 2 and 3 indicates that this phenomenon is fairly uncommon for these baseline models.

Given the high proportion of NULL-consensus examples in some languages, one concern is that a naïve baseline could simply predict “no answer” in all cases and achieve reasonable performance. It is therefore important to understand performance

Gemini Model →	1.5 Pro		2.0 Flash	
NULL consensus →	✓	✗	✓	✗
Language ↓				
Algerian Arabic	55.9	16.1	56.9	16.3
Egyptian Arabic	62.8	54.6	63.3	53.5
Iraqi Arabic	66.7	54.4	68.0	56.2
Jordanian Arabic	62.2	50.4	64.2	51.0
Armenian	81.6	40.3	81.1	39.4
Azerbaijani	83.0	60.9	81.4	61.0
Hebrew	64.6	50.3	63.5	49.4
Farsi	46.7	37.4	46.7	51.0
Tajik	71.9	58.3	62.5	44.6
Turkish	56.0	54.1	56.8	53.6

Table 5: Baseline F1 results on our test set, divided into examples with a NULL (✓) vs. non-NULL (✗) consensus.

on the non-NULL portion as well. Tables 4 and 5 present this breakdown, using 100-scaled F1 score[4]. Both baseline models achieve a higher performance on the NULL portion of both development and test splits in all languages except for Farsi. However, with the possible exception of Algerian Arabic, performance on the non-NULL data is still reasonable, indicating that the overall performance reported in Tables 2 and 3 is not a result of always predicting "no answer". We recommend that practitioners benchmarking performance on our dataset also report scores broken down in this way.

We emphasize that these baselines were not finetuned for this task, and that inference consists of showing the entire article to the model at once. Publicly-available models capable of achieving reasonable performance on such a long-context task are a recent development, and make it much easier for researchers to experiment on our task. As a point of contrast, the baseline for the original TyDi QA dataset (Clark et al., 2020) used a finetuned model that scored each candidate short answer span separately. We believe that the technological advances since then will make the MinSpan task easier for researchers to iterate on.

10 Conclusion

In this work we have presented a dataset of information-seeking questions in 10 low-resource language varieties, and demonstrated the feasibility of using modern LLMs to extract answers to these questions from large text contexts. By releasing the data and code to use it, we hope to facilitate the measurement and improvement of models' performance in these language varieties.

References

Joshua Ainslie, Santiago Ontanon, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. 2020. Etc: Encoding long and structured inputs in transformers. *arXiv preprint arXiv:2004.08483*.

Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. 2024. Calmqa: Exploring culturally specific long-form question answering across 23 languages. *arXiv preprint arXiv:2406.17761*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*.

Akari Asai, Jungo Kasai, Jonathan H Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2020. Xor qa: Cross-lingual open-retrieval question answering. *arXiv preprint arXiv:2010.11856*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. The madar shared task on arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207.

ElMehdi Boujou, Hamza Chataoui, Abdellah El Mekki, Saad Benjelloun, Ikram Chairi, and Ismail Berrada. 2021. An open access nlp dataset for arabic dialects: Data collection, labeling, and model construction. *arXiv preprint arXiv:2102.11000*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking

[4] For the NULL portion, F1 is equivalent to exact match.

question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Gulshan Dovudov, Vít Suchomel, and Pavel Šmerk. 2012. Pos annotated 50m corpus of tajik language. *Language Technology for Normalisation of Less-Resourced Languages*, page 93.

Leon Engländer, Hannah Sterz, Clifton Poth, Jonas Pfeiffer, Ilia Kuznetsov, and Iryna Gurevych. 2024. M2qa: Multi-domain multilingual question answering. *arXiv preprint arXiv:2407.01091*.

Gemini Team. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Gemini Team. 2024b. Gemini: A family of highly capable multimodal models.

Matthias Hecking and Tatiana Sarmina-Baneviciene. 2010. A tajik extension of the multilingual information extraction system zenon. In *Proceedings of the 15th International Command and Control Research and Technolgy Symposium (ICCRTS), Santa Monica, CA*. Citeseer.

Jafar Isbarov, Kavsar Huseynova, Elvin Mammadov, and Mammad Hajili. 2024. Open foundation models for azerbaijani language. *arXiv preprint arXiv:2407.02337*.

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.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Rayyan Merchant and Kevin Tang. 2024. Parstext: A digraphic corpus for tajik-farsi transliteration. In *Proceedings of the Second Workshop on Computation and Written Language (CAWL)@ LREC-COLING 2024*, pages 1–7.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jeanbaptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. 2020. Lareqa: Language-agnostic answer retrieval from a multilingual pool. *arXiv preprint arXiv:2004.05484*.

Sebastian Ruder, Jonathan H Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel A Sarr, Xinyi Wang, et al. 2023. Xtreme-up: A user-centric scarce-data benchmark for under-represented languages. *arXiv preprint arXiv:2305.11938*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297.

Jihad Zahir. 2022. Iadd: An integrated arabic dialect identification dataset. *Data in Brief*, 40:107777.

A Appendix

A.1 No-Answer Critic Prompt

Our no-answer critic used the following prompt:

> *You are given a text and you should check whether it indicates that an answer was not found or there is no answer, answer with no or yes only. Here are some examples.*
> *text: 25 AA response: No*
> *text: here AA response: No*
> *text: 2020 AA response: No*
> *text: YES AA response: No*
> *text: No AA response: No*
> *text: No Answer AA response:Yes*
> *text: there is no information in the text to answer the question AA response: Yes*
> *text: he was born in germany AA response: No*

AA is an arbitrary string separating the text and critic response. Other formatting can be used as well.