UQA: Corpus for Urdu Question Answering

Samee Arif, Sualeha Farid, Awais Athar, Agha Ali Raza
LUMS, LUMS, EMBL-EBI, LUMS
Lahore, Pakistan; Lahore, Pakistan; Hinxton, UK; Lahore, Pakistan
23100088@lums.edu.pk, 23100133@lums.edu.pk, awais@ebi.ac.uk, agha.ali.raza@lums.edu.pk

Abstract

This paper introduces UQA, a novel dataset for question answering and text comprehension in Urdu, a low-resource language with over 70 million native speakers. UQA is generated by translating the Stanford Question Answering Dataset (SQuAD2.0), a large-scale English QA dataset, using a technique called EATS (Enclose to Anchor, Translate, Seek), which preserves the answer spans in the translated context paragraphs. The paper describes the process of selecting and evaluating the best translation model among two candidates: Google Translator and Seamless M4T. The paper also benchmarks several state-of-the-art multilingual QA models on UQA, including mBERT, XLM-RoBERTa, and mT5, and reports promising results. For XLM-RoBERTa-XL, we have an **F1 score of 85.99 and 74.56 EM**. UQA is a valuable resource for developing and testing multilingual NLP systems for Urdu and for enhancing the cross-lingual transferability of existing models. Further, the paper demonstrates the effectiveness of EATS for creating high-quality datasets for other languages and domains. The UQA dataset and the code are publicly available at www.github.com/sameearif/UQA.

Keywords: Question-answering, machine translation, corpus, Urdu, low-resource languages, language resource, natural language processing

1. Introduction

The growth of natural language processing (NLP) tasks and datasets in English has been remarkable. However, expanding the reach of NLP to languages other than English, especially those that are lower on digital resources, is crucial for advancing multilingual AI systems. Among such languages, Urdu, with over 70 million native speakers[1], stands as a significant yet underrepresented language in the NLP domain.

The Stanford Question Answering Dataset 2.0 (SQuAD2.0) (Rajpurkar et al., 2018) is a benchmark for evaluating machine comprehension of text, but it is limited to English-based systems. There are two categories of questions: (1) Answerable questions: These are questions for which a clear, definite answer can be extracted directly from the provided passage or context (2) Unanswerable questions: These are questions for which the answer cannot be found in the provided passage but they look similar to answerable questions. Figure 1 shows examples of answerable and unanswerable question from SQuAD2.0.

Translating SQuAD2.0 into other languages seems like a straightforward task, but it comes with its own set of challenges, mainly when the job requires mapping the start index of the answer in the English context to the start index in the Urdu context. The introduction of the ”Enclose to Anchor, Translate, Seek” (EATS) technique addresses this very challenge by enclosing the answer within a context using a specific delimiter, translating the enclosed context, and then seeking the delimiter's position post-translation.

In this study, we contrast the outputs of popular translation models, including Google Translator[2] and Seamless M4T (Barrault et al., 2023). Through rigorous evaluation, we measure inter-rater agreement using Krippendorff's alpha (Krippendorff, 2004) to discern the most consistent and reliable translation method among the contenders. The selected model then serves as our primary tool in the EATS technique to produce the Urdu-translated dataset.

We intend for our work to serve as a tool to further the development of Urdu NLP tools to enable access to mainstream language applications among Urdu speakers. Due to the dataset's large size and high quality, it can serve as a valuable resource to train LLMs in Urdu and create domain-specific applications to empower underserved populations via educational and health resources.

2. Related Work

A large number of datasets for question-answering and text comprehension systems have been created for English. WikiQA (Yang et al., 2015) was introduced in 2015 - it included 3,047 questions and 29,258 sentences, where 1,473 sentences were

[1] www.britannica.com/topic/Urdu-language

[2] cloud.google.com/translate/docs/reference/rest

Paragraph:
The further decline of Byzantine state-of-affairs paved the road to a third attack in 1185, when a large Norman army invaded Dyrrachium, owing to the betrayal of high Byzantine officials. Some time later, Dyrrachium—one of the most important naval bases of the Adriatic—fell again to Byzantine hands.

Answerable Question:
When did the Normans attack Dyrrachium?
Answer: 1185

Unanswerable Question:
Who betrayed the Normans?

Figure 1: Question types example from SQuAD2.0

labeled as answer sentences to the questions. Soon after, Rajpurkar et al., 2016 introduced the Stanford Question Answering Dataset (SQuAD) created by crowdworkers posing questions on Wikipedia articles. They compiled 100,000+ questions for the task of machine comprehension of text. In an attempt to create more robust question-answering systems, a more challenging dataset titled SQuAD2.0 (Rajpurkar et al., 2018) was then introduced by expanding on the work done for SQuAD (Rajpurkar et al., 2016) - this introduced 50,000 unanswerable questions on top of the original dataset written adversarially by crowdworkers to look similar to answerable ones. Other corpora including HotpotQA (Yang et al., 2018) containing 113k Wikipedia-based question-answer pairs, TriviaQA (Joshi et al., 2017) with over 650K question-answer-evidence triples, and Meta's bAbI tasks data (Weston et al., 2016) were created to introduce greater complexity in data to train more capable QA systems.

Datasets for training cross-lingual functionality in QA systems were introduced in a multilingual context. The MLQA dataset was introduced by Lewis et al., 2019. It contains QA instances in SQuAD format in seven languages (English, Arabic, German, Spanish, Hindi, Vietnamese, and Simplified Chinese) and was built using an alignment strategy on Wikipedia articles. They generated over 12,000 instances in English and 5,000 instances in each other language. Similarly, XQuAD (Artetxe et al., 2019) (13,000 examples spanning 11 languages), XQA presented by Liu et al., 2019 (28,000 instances in 9 languages), TyDi by Clark et al., 2020 (204,000 examples in 11 languages), Xor-QA (Asai et al., 2021) (40,000 instances across seven languages), and MKQA (Longpre et al., 2021) (260,000 examples in 26 languages) was introduced for multilingual question answering systems.

There has been comparatively less work for monolingual non-English corpora - particularly for low-resource languages. A popular method of generating resources for such languages has been the translation of datasets for English into the target language employing different machine translation implementations. Some examples of such work are ParSQuAD (Abadani et al., 2021) for Persian, SQuAD-it (Croce et al., 2018) for Italian, Vietnamese SQuAD[3], K-QuAD[4] for Korean, and Arabic-SQuAD (Mozannar et al., 2019).

For Urdu, question-answering resources are scarce. Some datasets have been presented, such as UQuAD[5] containing 499 questions and 27 paragraphs. Urdu Open-Ended Question Answer Text Dataset[6] with 5000+ question-answer pairs, and Urdu Closed-Ended Question Answer Text Dataset[7] also containing 5,000+ question-answer pairs are both human-generated datasets however they are not open source and do not provide any metrics or comments regarding the quality of the data. UQuAD1.0 (Kazi and Khoja, 2021) is a work involving the translation of SQuAD1.0 (Rajpurkar et al., 2016) containing 49,000 question-answer pairs from which 45,000 are translated from SQuAD (Rajpurkar et al., 2016) (53% of the data, the remaining 47% was discarded) and 4,000 were manually generated via crowdsourcing. However, the dataset is not publicly available. Therefore, to the best of our knowledge, no large, high-quality, publicly available dataset exists for Urdu question answering and text comprehension - making our contribution a valuable and important step towards developing tools for this low-resource language.

3. Methodology

Translating the Stanford Question Answering Dataset (SQuAD2.0) (Rajpurkar et al., 2018) into Urdu presents a unique set of challenges, with one of the foremost difficulties lying in accurately identi-

[3] www.kaggle.com/datasets/nkhachao/vietnamese-squad

[4] www.github.com/Di-lab-Yonsei/K-QuAD

[5] www.github.com/ahsanfarooqui/UQuAD---Urdu-Question-Answer-Dataset/tree/main

[6] www.futurebeeai.com/dataset/prompt-response-dataset/urdu-open-ended-question-answer-text-dataset

[7] www.futurebeeai.com/dataset/prompt-response-dataset/urdu-closed-ended-question-answer-text-dataset

fying the answer's starting position within the translated context paragraph. This challenge stems from the linguistic differences between the source and target language. Both languages have different grammatical structures, vocabulary, and idiomatic expressions, meaning there is no one-to-one mapping between the words in the source text and the translated text. The source language and the target language also have different word order and sentence structure, that is, English follows subject-verb-object (SVO) order, and Urdu follows subject-object-verb (SOV) order. Therefore, addressing this challenge requires a robust method for aligning and matching the answer spans.

3.1. Translation Model Selection

All SQuAD2.0 (Rajpurkar et al., 2018) context paragraphs were split into sentences using the python NLTK (Bird and Loper, 2004) sentence tokenizer for our experiments. For experiment 1, we selected a set of 100 sentences to conduct a pilot test. This smaller subset allowed us to assess the viability of our translation methodology and the overall experimental design in a controlled, manageable environment. In experiment 2 we selected a set of 1,512 sentences from a total of 100,026 sentences. The minimum required sample size for a confidence level of 99% with a 3% margin of error was calculated to be 1,030 using the population size (100,000) - we therefore took a sample of 1,512 sentences. The selected sentences for both experiments were subsequently passed through two machine translation systems: Google Translator[2] and Facebook Seamless M4T (Barrault et al., 2023).

In evaluation 1, three annotators (computer science researchers - native Urdu speakers) were presented with two anonymized machine translation systems, one of which was the Seamless M4T model and the other was Google Translator. The annotators were each assigned a total of 100 identical sentences. Their task involved labeling the data to indicate one of the following: (1) both translators produced the same output quality for a sentence, (2) Seamless M4T provided a better translation, or (3) Google Translator provided a better translation. The Google Translator was picked 14.33% of the times, and Seamless M4T was picked 51.67% of times, and both were considered to be of the same quality 34.0% of times. To determine the inter-rater reliability, the Krippendorff's alpha value was calculated and found to be 0.688, which, according to Krippendorff's interpretation (Krippendorff, 2004) is sufficient for a tentative conclusions to be drawn.

In evaluation 2, twelve voters - undergraduate students, native Urdu speakers with English as medium of instruction - were asked to pick between the two translation models. The voters were given the same task as the annotators in Experiment 1. The Google Translator was picked 37.43% of the times, and Seamless M4T was picked 54.37% of the times, and both were considered to be of the same quality 8.20% of times.

In summary, Seamless M4T consistently demonstrated superior translation quality in both evaluations when compared with Google Translator.

3.2. Initial Experiments

Our initial approach involved translating the SQuAD2.0 dataset (Rajpurkar et al., 2018) by translating the question and the answer string and then translating the context sentence by sentence. When we reached the line where the answer string was present, we used string matching to find the translated answer. However, we encountered significant challenges: (1) The answer string in the context often underwent grammatical modifications when included within a paragraph; (2) The sentence tokenization libraries failed to detect all the abbreviations, resulting in low-quality sentence segmentation and, therefore, degraded translation results. As a result, relying solely on the exact string-matching approach proved to be insufficient for pinpointing the answer's start index.

In our second approach, we opted to translate the each undivided paragraph as a single unit (to retain the context and semantic meaning) instead of translating line by line. This shift in strategy allowed us to pass the translated paragraph alongside the translated question and answer to a Large Language Model (LLM), such as LLaMA 2 (Touvron et al., 2023) and GPT-3.5 (Brown et al., 2020), in an attempt to determine the answer's start and end positions automatically. However, the models performed poorly on our text as they did not predict the correct start and end points in the Urdu paragraphs.

Subsequently, we transitioned to using GPT-4 (OpenAI, 2023), demonstrating promising results in accurately identifying answer positions within the translated context paragraph. However, the drawback of this approach was the significant computational cost associated with GPT-4, which rendered it impractical for this task. As a result, we had to reconsider our methodology to balance performance with computational efficiency and cost.

To address the challenges encountered in translating the SQuAD2.0 dataset into Urdu and accurately identifying answer start positions, we implemented a three-step solution illustrated in Figure 2.

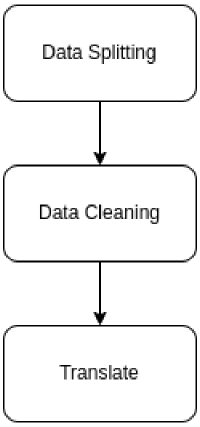

Figure 2: Three-step solution

3.3. Implementation

3.3.1. Data Splitting

When a large paragraph (defined as text containing more than 1,000 characters) is passed through Seamless M4T (Barrault et al., 2023), it tends to summarize or drop the last few sentences. To illustrate this behavior, Table 1 presents sentences from a paragraph of length 1,427 characters. The entire paragraph is passed through the translator, then the English paragraph is manually split into sentences, and the corresponding sentences are extracted from the translated paragraph. It is evident from the output that the last sentence is not fully translated.

To address this issue, we initially identified paragraphs with a length equal to or exceeding 1,000 characters. We then manually divided these 3,307 paragraphs into smaller segments, ensuring that each paragraph segment had a length of less than 1,000 characters.

3.3.2. Data Cleaning - EATS

To ensure that answer strings are retained in the translated text and unaffected by any text misalignment during the translation process, we introduce the EATS technique: Enclose to Anchor, Translate, Seek. The process involves first highlighting the answer string in the original text by enclosing it in delimiters, then passing the text through a translator and seeking the answers in the target language by looking for the delimiters. Thus, the first part of our process was to ensure that the data was in the following format:

Infrared radiation is used in industrial, ••scientific•• and medical applications.

The answer was marked with the delimiters because the removal of certain characters from the string would offset the answer start position, so '••' acts as a marker for the answer string in the context paragraph.

English	Urdu Translation
The Norman dynasty had a major political, cultural and military impact on medieval Europe and even the Near East.	نارمن خاندان نے قرون وسطی کے یورپ اور یہاں تک کہ مشرق وسطی پر ایک بڑا سیاسی ، ثقافتی اور فوجی اثر ڈالا۔
The Normans were famed for their martial spirit and eventually for their Christian piety, becoming exponents of the Catholic orthodoxy into which they assimilated.	نارمن اپنے جنگی جذبے اور بالآخر اپنی عیسائی عقیدت کے لئے مشہور تھے ، جو کیتھولک آرتھوڈوکس کے نمائندے بن گئے جس میں انہوں نے ہم آہنگ کیا۔
…	
Norman cultural and military influence spread from these new European centres to the Crusader states of the Near East, where their prince Bohemond I founded the Principality of Antioch in the Levant, to Scotland and Wales in Great Britain, to Ireland, and to the coasts of north Africa and the Canary Islands.	اسکاٹ لینڈ ، آئرلینڈ اور شمالی افریقہ میں ان کے ثقافتی مراکز تک پھیل گئے۔

Table 1: English paragraph to Urdu translation

Seamless M4T (Barrault et al., 2023) sometimes fails to handle semicolons ';', en dashes '–' that are used between figures to represent the range and em dashes '—' that are used to create a strong break in a sentence, emphasizing an interruption or additional information. To account for this, we replaced all the semicolons with the Arabic semicolon '؛' and en dashes and em dashes with double hyphens '−−' before the translation process started. Following this, all double quotation marks were removed from the text to ensure that only the answer string is enclosed within the specific double quotation mark '"' (i.e. U+0022 in UTF-16 encoding) for the translation process. In the final step '••' was replaced with '"'. This data cleaning procedure was carried out for all the paragraphs, questions, and answers in SQuAD2.0 (Rajpurkar et al., 2018).

3.3.3. Translation

As we had used quotation marks to highlight the answer paragraphs, they would serve as essential markers to identify the answer's start and end positions within the translated paragraph and precisely locate the answer within the text. Algorithm 1 outlines the pseudo-code for this process. In the algorithm, variable $paragraphs$ is a list containing either a single entity if it contains less than or equal to 1000 characters, otherwise it contains multiple entities i.e. sub-paragraphs - this is due to the splitting methodology defined in section 3.3.1. Following the translation, the double hyphens '$--$' in between digits were replaced with en dash '–', and the rest (i.e those not between digits) were replaced with em dash '—'.

Algorithm 1 Translation Algorithm

```
for $([paragraphs], question, answer)$ in $data$ do
    $question_t \leftarrow translate(question)$
    $answer_t \leftarrow translate(answer)$
    $paragraph_t \leftarrow [\,]$
    for $para$ in $paragraphs$ do
        if '••' is in $para$ then
            $Replace$ '••' $with$ '"' $in\ para$
            $para \leftarrow translate(para)$
            if $para.count('"') \neq 2$ then
                break
            end if
        else
            $para \leftarrow translate(para)$
        end if
        $paragrapht_t.append(para)$
    end for
end for
```

Our methodology failed to retain quotation marks for only 392 out of 11,858 questions in the dev set and 5,574 out of 130,319 questions in the train set. Therefore, from a total of 142,177 questions only 5,966 were discarded which highlights the effectiveness of our approach, demonstrating a high degree of precision. In addressing the issue of missing quotation marks in a minor subset of our dataset, we found that GPT-4 (OpenAI, 2023) could effectively correct these errors. However, considering that the erroneous data was only 4.2% of the overall dataset, we ultimately decided against correcting this subset using GPT-4 with the aim of minimizing expenses.

Following the translation and implementation of EATS, we generated a total of 124,745 questions in the train set and 11,466 questions in the dev set. A breakdown of the number of questions in each category is provided in Table 2.

	Dev	**Train**
Answerable Questions	5,811	83,018
Unanswerable Questions	5,655	41,727

Table 2: Dataset summary

4. Evaluation and Discussion

We fine-tuned and evaluated different variants of mBERT (Devlin et al., 2019), XLM-RoBERTa (abbreviated as XLM-R) (Conneau et al., 2020; Goyal et al., 2021) and mT5 (Xue et al., 2021) on our dataset. All the models were fine-tuned for 4 epochs, learning rate for XLM-R models and mBERT was set to $2e^{-5}$ and $5e^{-5}$ for mT5 models. We used only answerable questions for fine-tuning on the train set and evaluating these models on the dev set. The performance of the models is quantified using two common metrics: Exact Match (EM) and F1 Score using the Huggingface wrapper[8] for the official SQuAD evaluation script by Rajpurkar et al., 2016. All the models were trained for six epochs and the best checkpoint of each model was evaluated on the dev part of the dataset. Table 3 summarizes the results of the experiments.

Model	**F1 Score**	**Exact Match (EM)**
mBERT	64.72	45.50
mT5-Small	67.24	52.37
mT5-Large	84.20	71.26
XLM-R	78.00	65.67
XLM-R-Large	84.42	72.24
XLM-R-XL	**85.99**	**74.56**

Table 3: Evaluation summary

We can see that the XLM-R-XL performs the best for both metrics, with mT5-Large closely following. This can be explained by the number of parameters (3.5B vs 1.2B) as well as differences in the size and quality of their original training corpora.

Comparing the results with existing state-of-the-art models for Urdu and similar languages (Persian and Arabic), Table 4 shows that our XLM-R-XL UQA model outperforms the best reported scores. While these results are not directly compa-

[8] www.huggingface.co/spaces/evaluate-metric/squad

rable with existing work due to differences in model parameter sizes, results of our evaluation on comparable models including mBERT and XLM-R (table 3) show that models trained on UQA outperform those presented for Arabic-SQuAD (BERT), and UQuAD1.0 (XLM-R). This improvement can be attributed to the quality of translation as well as the size of our training data.

Incorporating unanswerable questions from our dataset into the training set could present a valuable opportunity to enhance model performance. Training on both answerable and unanswerable questions might empower the model to better discern between the two, potentially refining its ability to identify and respond to answerable queries with increased precision.

Dataset	**Model**	**F1 Score**	**Exact Match (EM)**
ParSQuAD	ALBERT	70.84	67.73
Arabic-SQuAD	BERT	61.30	34.20
UQuAD1.0	XLM-R	66.00	36.00
UQA	**XLM-R-XL**	**85.99**	**74.56**

Table 4: Comparison with existing models

5. Conclusion

In this paper, we present the process of creating a question answering corpus for Urdu and make UQA publicly available. By training multiple state of the art question answering models on our datasets to get promising evaluation scores, we demonstrate the suitability of our dataset for training and evaluation of transformer based models. Future work can include building on to the dataset with domain specific data to fine-tune models - particularly LLMs - for a specific use case such as providing health care facilities to low resource areas.

In the translation process, we primarily relied on the selected model's inherent accuracy, given that no translation model guarantees 100% accuracy. We also did an extensive evaluation of the translation models to ensure that the one with the highest accuracy was used. In a low-resource language like Urdu achieving perfect translation accuracy can be a challenge, the large size of the dataset also makes manual fixes infeasible.

This paper also forms the groundwork for a pipeline to produce further domain-specific QA resources for Urdu without the need for translation by relying directly on question generation models that can be trained on UQA.

Our work for resource generation in low resource languages, therefore, creates the opportunity to address the challenge of large-scale data generation required for language models across diverse languages and domains. Particularly in contexts where native data in the target language is sparse or unavailable.

6. Acknowledgements

We are grateful for the time and effort put in by our research interns who annotated our data: Aamina Jamal Khan, Abdullah Hashmat, Abeer Aamir, Daanish Uddin Khan, Eesha Kamran, Maham Ahmed Bhimra, Maryam Usman, Muhammad Faisal Nazir, Muhammad Khaquan, Muhammad Shayan, Omer Tafveez, and Qasim Anwar who supported our work.

7. Bibliographical References

Negin Abadani, Jamshid Mozafari, Afsaneh Fatemi, Mohammd Ali Nematbakhsh, and Arefeh Kazemi. 2021. Parsquad: machine translated squad dataset for persian question answering. In *2021 7th International Conference on Web Research (ICWR)*, pages 163–168. IEEE.

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 Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564, Online. Association for Computational Linguistics.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2023. Seamlessm4t-massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

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

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Danilo Croce, Alexandra Zelenanska, and Roberto Basili. 2018. Neural learning for question answering in italian. In *AI*IA 2018 – Advances in Artificial Intelligence*, pages 389–402, Cham. Springer International Publishing.

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.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Samreen Kazi and Shakeel Khoja. 2021. Uquad1. 0: Development of an urdu question answering training data for machine reading comprehension. *arXiv preprint arXiv:2111.01543*.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*, 2nd edition. Sage Publications.

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

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, Florence, Italy. Association for Computational Linguistics.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering.

Hussein Mozannar, Elie Maamary, Karl El Hajal, and Hazem Hajj. 2019. Neural Arabic question answering. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 108–118, Florence, Italy. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh

Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomás Mikolov. 2016. Towards ai-complete question answering: A set of prerequisite toy tasks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

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.

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, Lisbon, Portugal. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.