SynDARin: Synthesising Datasets for Automated Reasoning in Low-Resource Languages

Gayane Ghazaryan[†1] **Erik Arakelyan**[†2] **Pasquale Minervini**[3] **Isabelle Augenstein**[2]
[1]American University of Armenia [2]University of Copenhagen
[3]University of Edinburgh
gayane_ghazaryan2@edu.aua.am erik.a@di.ku.dk
p.minervini@ed.ac.uk augenstein@di.ku.dk

Abstract

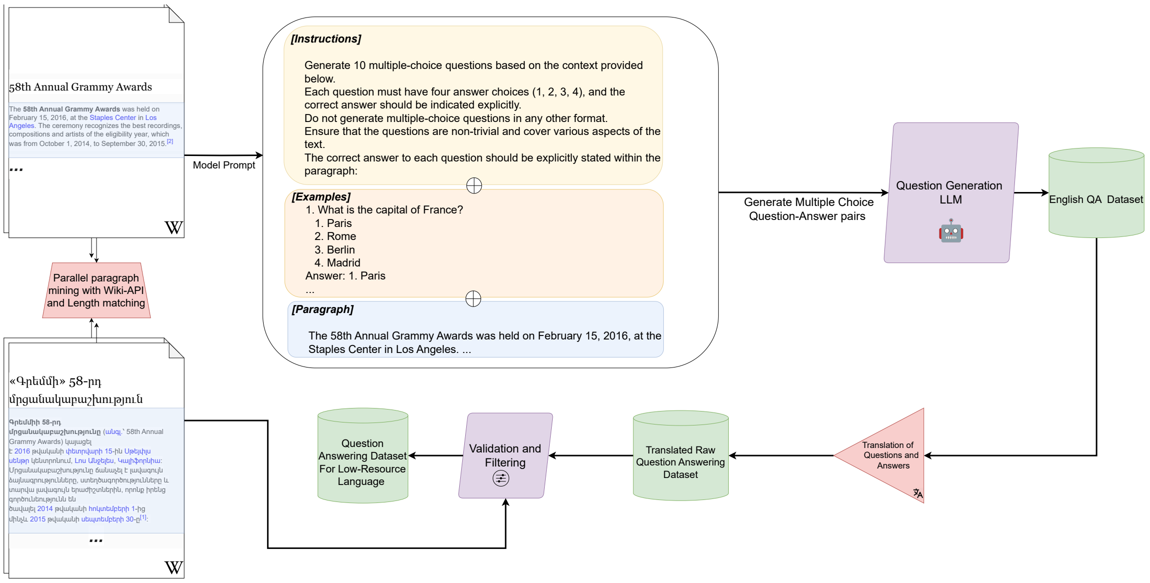
Question Answering (QA) datasets have been instrumental in developing and evaluating Large Language Model (LLM) capabilities. However, such datasets are scarce for languages other than English due to the cost and difficulties of collection and manual annotation. This means that producing novel models and measuring the performance of multilingual LLMs in low-resource languages is challenging. To mitigate this, we propose **Syn**DAR**in**, a method for generating and validating QA datasets for low-resoucre languages. We utilize parallel content mining to obtain *human-curated* paragraphs between English and the target language. We use the English data as context to *generate* synthetic multiple-choice (MC) question-answer pairs, which are automatically translated and further validated for quality. Combining these with their designated non-English *human-curated* paragraphs form the final QA dataset. The method allows to maintain content quality, reduces the likelihood of factual errors, and circumvents the need for costly annotation. To test the method, we created a QA dataset with 1.2K samples for the Armenian language. The human evaluation shows that 98% of the generated English data maintains quality and diversity in the question types and topics, while the translation validation pipeline can filter out $\sim 70\%$ of data with poor quality. We use the dataset to benchmark state-of-the-art LLMs, showing their inability to achieve human accuracy with some model performances closer to random chance. This shows that the generated dataset is non-trivial and can be used to evaluate reasoning capabilities in low-resource language.

1 Introduction

Question Answering (QA) has been a hallmark task for testing reading comprehension and reasoning capabilities in NLP systems. The availability of numerous English benchmarks that frame the problem as extractive, cloze-style or open-domain (Yang et al., 2015; Rajpurkar et al., 2016; Chen et al., 2017) reasoning tasks, along with novel pretrained language models (PLMs) (Devlin et al., 2018; Lewis et al., 2019a) and LLMs (Touvron et al., 2023; Jiang et al., 2023; Achiam et al., 2023) allowed for the development and granular evaluation of QA systems that occasionally boast human-like or better performance (Devlin et al., 2018; Min et al., 2023; Rogers et al., 2023). Although some concentrated effort has been made to create multilingual QA resources (Lewis et al., 2019b; Asai et al., 2018; Liu et al., 2019), the datasets remain rather scarce and usually cover a small selected set of languages due to the labour-intensive annotation costs. The proposed methods suggest using direct machine translation (Lewis et al., 2019b; Carrino et al., 2019) or multilingual synthetic data generation (Riabi et al., 2020; Agrawal et al., 2023; Shakeri et al., 2020). However, these approaches are directly bound to introduce biases and hallucinations during translation (Artetxe et al., 2020), cross-lingual transfer (Lauscher et al., 2020; Guerreiro et al., 2023) or generation (Ahuja et al., 2023). These limitations directly hinder the possibility to *develop* and *evaluate* the multilingual QA capabilities of language models in low-resource languages.

In this work, we propose **Syn**DAR**in**, a novel method for synthesising datasets for automated reasoning in low-resource languages that circumvents the above-mentioned obstacles and test it by creating a QA dataset for the Armenian language, which has virtually no presence of structured NLP datasets (Avetisyan and Broneske, 2023). We mine parallel English and Armenian introductory paragraphs from the same diverse set of Wikipedia articles, ensuring that the contents match by comparing their relative length. Similar mining approaches have been shown to be efficient for this task (Lewis

[†]Equal contribution



58th Annual Grammy Awards
...
W
Model Prompt
Parallel paragraph mining with Wiki-API and Length matching
«Գրեմմի» 58-րդ մրցանակաբաշխություն
...
W
Generate Multiple Choice Question-Answer pairs
Question Generation LLM
English QA Dataset
Translation of Questions and Answers
Translated Raw Question Answering Dataset
Validation and Filtering
Question Answering Dataset For Low-Resource Language

The **58th Annual Grammy Awards** was held on February 15, 2016, at the Staples Center in Los Angeles. The ceremony recognizes the best recordings, compositions and artists of the eligibility year, which was from October 1, 2014, to September 30, 2015.[2]

Գրեմմիի 58-րդ մրցանակաբաշխությունը (անգլ.՝ 58th Annual Grammy Awards) կայացել է 2016 թվականի փետրվարի 15-ին Սթեյփլս սենթր կենտրոնում, Լոս Անջելես, Կալիֆորնիա: Մրցանակաբաշխությունը ճանաչել է լավագույն ձայնագրությունները, ստեղծագործությունները և տարվա լավագույն երաժիշտներին, որոնք իրենց գործունեությունն են ծավալել 2014 թվականի հոկտեմբերի 1-ից մինչև 2015 թվականի սեպտեմբերի 30-ը[1]:

[Instructions]

Generate 10 multiple-choice questions based on the context provided below.
Each question must have four answer choices (1, 2, 3, 4), and the correct answer should be indicated explicitly.
Do not generate multiple-choice questions in any other format.
Ensure that the questions are non-trivial and cover various aspects of the text.
The correct answer to each question should be explicitly stated within the paragraph:

[Examples]

1. What is the capital of France?
   1. Paris
   2. Rome
   3. Berlin
   4. Madrid

Answer: 1. Paris
...

[Paragraph]

The 58th Annual Grammy Awards was held on February 15, 2016, at the Staples Center in Los Angeles. ...

Figure 1: The proposed framework is comprised of three components: (i) a module for mining parallel paragraphs using wiki-API and length matching; (ii) generating a synthetic question-answering dataset with an LLM using the mined English paragraphs; (iii) translating the question-answer pairs and Filtering/Validating them for obtaining a high-quality synthetic QA dataset in the low-resource language.

et al., 2021; Artetxe and Schwenk, 2019). This allows us to obtain human-curated text from diverse topics while bypassing a wide chunk of direct content translation and annotation. Given the English subset of this data, we generate MC question-answer pairs by prompting an LLM to produce queries with an answer explicitly mentioned within the paragraph. Following Lewis et al. (2019b), we filter out examples that do not contain the answer substring verbatim in the paragraph and additionally perform a human evaluation on a subset of 50 examples and show that 98% of these question-answer pairs are answerable and maintain quality. The produced question-answers are subsequently translated using an automated tool and further validated by answer substring and semantic matching in the parallel Armenian paragraph. This allows us to mitigate the likelihood of hallucinated, biased and inconsistent entries in the final QA dataset. Our human evaluation with native Armenian speakers shows that 70% of such corrupted examples are removed. We use the dataset as a reasoning benchmark for Armenian and evaluate several LLMs in zero-shot, few-shot, and fine-tuned modes. We show that the dataset cannot be trivially solved, thus highlighting it as a useful resource for measuring model performance. In sum, our contributions are as follows: (i) a novel method for QA dataset construction in low-resource languages, (ii) a QA dataset in Armenian, (iii) ablations showing the quality of the generated samples and (iv) an evaluation of several LLM families on the QA dataset.

2 Methodology

An outline of Syn**DAR**in can be seen in Fig. 1.

2.1 Parallel Data Mining

Given parallel English and Armenian introductory paragraph tokens $\mathcal{P}_{\text{En}} = (T_1, \dots T_n)$, $\mathcal{P}_{\text{Arm}} = (T_1, \dots T_m)$ obtained from a diverse set of Wiki articles, we want to save the segments that contain the same content. As the introductory paragraphs in Wikipedia contain highly similar information (Lewis et al., 2019b), we found that filtering out the paragraph pairs based on their relative view count and the number of tokens, i.e. length, is sufficient. To do this, we simply define a conditional rejection process on Wikipedia pages that have been viewed more than 1000 and edited more than 5 times $|||\mathcal{P}_{\text{En}}|| - ||\mathcal{P}_{\text{Arm}}||| \leq K_{\text{DM}}$, where K_{DM} is the threshold for the length difference. A higher length difference would imply that the contents of the paragraphs are misaligned, thus making us reject such samples. Consequently, we are able to obtain naturally written human-curated parallel paragraphs that cover a diverse set of topics.

2.2 QA Generation

After obtaining the parallel data, we prompt an LLM $\mathcal{M}$ with instructions $\mathcal{I} = (T_1, \dots T_{|\mathcal{I}|})$

Who	Where	What	When	Which	How	General	Why
304	128	1536	215	473	244	76	16

Table 1: Frquency of Question Types in the generated English question-answer pairs.

and 10 in-context example demonstrations $\mathcal{E} = (E_1, \dots E_{10})$, where $\forall i, E_i = (T_1, \dots T_{|E_i|})$, to generate diverse English MC question-answer pairs $\mathcal{K}_{\text{Eng}} = \{(q_1, a_1) \dots (q_N, a_N)\}$ given an English context paragraph $\mathcal{P}_{\text{En}}$:

$$q_i, a_i \sim \prod_{t=1}^{|\mathcal{K}_i|} P_{\mathcal{M}} \left(T_t^{(i)} \mid T_1^{(i)}, ..., T_{t-1}^{(i)}, \mathcal{I}, \mathcal{E}, \mathcal{P}_{\text{En}} \right) \quad (1)$$

We filter out all repeating questions, $\forall \{i, j : i \neq j\}, q_i \neq q_j$, and question-answers pairs where the answer span is not exactly mentioned within the text, i.e. $a_i \not\subset \mathcal{P}_{\text{En}}$. An example input used for generation can be seen in Fig. 1. This generation and validation pipeline resembles the ones in Lewis et al. (2021); Agrawal et al. (2023), which have shown successful question-generation results for the English language. Several examples of produced questions are available in Appendix A.

2.3 Translation and Validation

We transfer the generated question-answer pairs $\mathcal{K}_{\text{Eng}}$ into Armenian by using the Google Translate API to obtain $\mathcal{K}_{\text{Arm}}$. To mitigate the inconsistencies introduced during the translation process, we save only the samples where the translated answer $a_i \in \mathcal{K}_{\text{Arm}}$ is contained within and semantically related to the paragraph $\mathcal{P}_{\text{Arm}}$. To do this, we use a fuzzy substring matching function $\mathcal{F} : \mathcal{T} \times \mathcal{T} \rightarrow [0, 1]$, along with a multilingual language model $\mathcal{M}_{\text{sim}} : \mathcal{T} \rightarrow \mathcal{R}^d$ to measure semantic similarity, where $\mathcal{T}$ is an arbitrary set of tokens and d is the dimensionality of the embedding space of the model. Samples below a certain threshold, $\mathcal{F}(a_i, \mathcal{P}_{\text{Arm}}) \leq K_{\text{Fuzz}}$ and $\cos(\mathcal{M}(a_i), \mathcal{M}(\mathcal{P}_{\text{Arm}})) \leq K_{\text{Sim}}$ are filtered out. Note that exact matching is insufficient, as the morphology of the translated answer tokens can vary in the low-resource language. The multiple-choice answers are balanced uniformly in the final dataset so as not to introduce a bias toward any particular answer ordering.

3 Experimental Setup

QA Generation Our QA generation uses GPT-4 (Achiam et al., 2023), known for generating high-quality text (Zhou et al., 2023) and synthetic data (Hämäläinen et al., 2023; Li et al., 2023).

Problem type(%)	*Filtered*	*Unfiltered*
Partially Missing Info	38	77
Bad Translation	5	51
Partially Correct Answers	22	31
Several Correct Answers	27	45
Date Mismatch	13	17
Other	8	22

Table 2: Unanswerable sample analysis before(Unfiltered) and after(Filtered) the validation. Annotators can choose multiple reasons per sample.

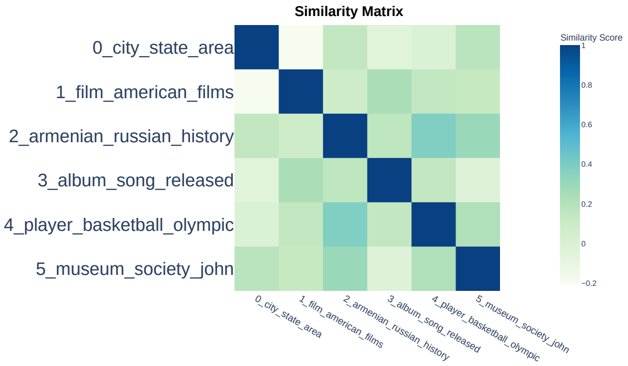


Figure 2: BERTopic embeddings similarity heatmap for the top 6 frequent topics in the mined English paragraphs.

Substring Matching and Semantic Similarity We employ Levenshtein distance for fuzzy substring matching ($\mathcal{F}$) and multilingual sentence embeddings (Reimers and Gurevych, 2019) ($\mathcal{M}_{\text{sim}}$) for semantic similarity using cosine distance.

Armenian QA Benchmarking We benchmark GPT-3.5 (Achiam et al., 2023), CMD-R, and CMD-R+ (Cohere, 2024) using $\{0, 2, 4, 6\}$ in-context examples with few-shot prompting (Brown et al., 2020) on the Armenian QA dataset. We further frame the task as classification with multiple-choice answers and perform supervised fine-tuning with a recipe (Mosbach et al., 2021) on XLM-RoBERTa-base (Conneau et al., 2019), with $\{32, 64, \dots, 980\}$ training samples and benchmark it on the same testing set. Following Poliak et al. (2018), we analyze model performance on *question-only* and *paragraph-only* inputs for bias detection.

4 Results

4.1 English QA Dataset Generation

We mined 300 parallel English-Armenian Wikipedia paragraphs and generated 10 diverse

	Accuracy			
Filter	128	256	512	987
Complete	30.1%	33.5%	38.7%	39.5%
paragraph-only	26.7%	28.3%	23.9%	28.3%
question-only	22.1%	22.7%	19.4%	23.5%
Random performance	**25.0%**			

Table 3: The results of fine-tuning XLM-Roberta on the Armenian QA dataset with a varying number of training samples in different degeneracy testing scenarios.

	Accuracy			
Model Name	0	2	4	6
Command-R	58.7%	**68.4%**	64.8%	64.0%
Command-R+	59.3%	67.2%	69.6%	**70.9%**
GPT-3.5	56.3%	56.3%	**59.1%**	54.3%

Table 4: Model Accuracy with a varying number of provided in-context samples before generation.

questions with 4 MC answers each, resulting in 3000 English QA pairs.

Dataset Diversity We assessed question diversity (Table 1) and found meaningful variation consistent with prior human-curated datasets (Lewis et al., 2019b; Rajpurkar et al., 2016). Topic modelling using BERTopic (Grootendorst, 2022) validated the subject diversity (Fig. 2). A granular diversity analysis within the dataset is presented in Appendix A.

Human Evaluation To assess the data quality, we follow Lewis et al. (2021) and ask two English-speaking human annotators to manually inspect 50 randomly chosen samples from the English QA dataset regarding the captured contextual information and answerability of the sample question. The results show, with an inter-annotator agreement score of Cohen's $\kappa = 0.99$, that 98% of examples contain sufficient details to answer the question while accurately capturing contextual information.

4.2 Automatic Translation and Validation

We translate the obtained 3000 QA samples and pass the results through our validation pipeline to produce 1235 filtered Armenian examples.

Armenian QA dataset We use these samples and their designated Armenian paragraphs to form the QA dataset. We split the data into 80/20 *train/test* buckets with 987 samples in training and 247 in testing. We ensure that the paragraphs in the testing set are not contained in the train set to avoid any data leakage. We maintain a uniform distribution of MC questions within the answers, avoiding bias towards any answer ordering.

Human Evaluation We assessed the translation validation pipeline and datasets using two native-speaking annotators. They reviewed the *test* set, which was mixed with 100 randomly flagged poor samples from automatic validation. Annotators either answered the samples or marked them as unanswerable, citing reasons from a predefined set, see in Table 2. Results showed that 87% of the flagged examples were unanswerable due to insufficient context, translation errors, or hallucinations. The error breakdown in Table 2 highlights the quality improvement in filtered samples w.r.t. to the abovementioned discrepancies, where annotators answered correctly in 75% of cases. We measure the inter-annotator agreement using Cohen's $\kappa = 0.8$. These confirm the ability of our validation pipeline to maintain the dataset quality.

Benchmarks To show the value of the created dataset, we investigate if it suffers from statistical biases or degenerate solutions by training an XLM-RoBERTa model on inputs that contain only the paragraph or the question, excluding everything else from the sample. The results in Table 3 show that regardless of the number of training samples, the models trained with question and paragraph-only samples behave similarly to random chance, while training with complete data gradually increases the performance, highlighting that the dataset is unlikely to suffer from inconsistencies and degenerate solutions and can be used for developing QA capabilities for Armenian. We further benchmark several state-of-the-art LLMs on this dataset in supervised fine-tuning, *zero-shot* and *few-shot* settings. We see in Table 4 that even the largest models do not trivially solve the dataset, showing its utility as a benchmarking tool.

5 Conclusion

We propose **SynDARin**, a novel method for constructing QA datasets for low-resource languages and produce a dataset for the Armenian language. Systematic studies of the reliability of the individual modules to produce diverse QA samples that maintain answerability and quality show the effectiveness of the method. We further use the produced Armenian QA dataset to benchmark state-of-the-art LLMs and show the value of the proposed resource in evaluating QA reasoning capabilities in the low-resource language.

Limitations

The proposed methods have currently been tested only for a smaller-scale QA dataset creation in Armenian, thus not allowing us to complete a wider cross-lingual study. The study benchmarks should be extended and analyzed further in more multilingual, low-resource languages. In the case of extremely rare low-resource languages, the automatic translation part within our pipeline would require either the development of such a translation method, robust cross-lingual transfer from a similar language or direct manual effort, all of which are bound to introduce either qualitative or logistic complications while creating the final QA resource.

Acknowledgments

Erik is partially funded by a DFF Sapere Aude research leader grant under grant agreement No 0171-00034B, as well as by a NEC PhD fellowship, and is supported by the Pioneer Centre for AI, DNRF grant number P1. Pasquale was partially funded by ELIAI (The Edinburgh Laboratory for Integrated Artificial Intelligence), EPSRC (grant no. EP/W002876/1), an industry grant from Cisco, and a donation from Accenture LLP. Isabelle's research is partially funded by the European Union (ERC, ExplainYourself, 101077481), and is supported by the Pioneer Centre for AI, DNRF grant number P1. This work was supported by the Edinburgh International Data Facility (EIDF) and the Data-Driven Innovation Programme at the University of Edinburgh.

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*.

Priyanka Agrawal, Chris Alberti, Fantine Huot, Joshua Maynez, Ji Ma, Sebastian Ruder, Kuzman Ganchev, Dipanjan Das, and Mirella Lapata. 2023. Qameleon: Multilingual qa with only 5 examples. *Transactions of the Association for Computational Linguistics*, 11:1754–1771.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. *arXiv preprint arXiv:2004.04721*.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the association for computational linguistics*, 7:597–610.

Akari Asai, Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2018. Multilingual extractive reading comprehension by runtime machine translation. *arXiv preprint arXiv:1809.03275*.

Hayastan Avetisyan and David Broneske. 2023. Large language models and low-resource languages: An examination of armenian nlp. *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 199–210.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Casimiro Pio Carrino, Marta R Costa-Jussà, and José AR Fonollosa. 2019. Automatic spanish translation of the squad dataset for multilingual question answering. *arXiv preprint arXiv:1912.05200*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Cohere. 2024. Command r: Retrieval-augmented generation at production scale. `https://txt.cohere.com/command-r`.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Nuno M Guerreiro, Duarte M Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python. If you use spaCy, please cite it as below.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. *arXiv preprint arXiv:2005.00633*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019a. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

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

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. Paq: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. *arXiv preprint arXiv:2310.07849*.

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.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning BERT: misconceptions, explanations, and strong baselines. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042*.

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

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

Arij Riabi, Thomas Scialom, Rachel Keraron, Benoît Sagot, Djamé Seddah, and Jacopo Staiano. 2020. Synthetic data augmentation for zero-shot cross-lingual question answering. *arXiv preprint arXiv:2010.12643*.

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

Siamak Shakeri, Noah Constant, Mihir Sanjay Kale, and Linting Xue. 2020. Towards zero-shot multilingual synthetic question and answer generation for cross-lingual reading comprehension. *arXiv preprint arXiv:2010.12008*.

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

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.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

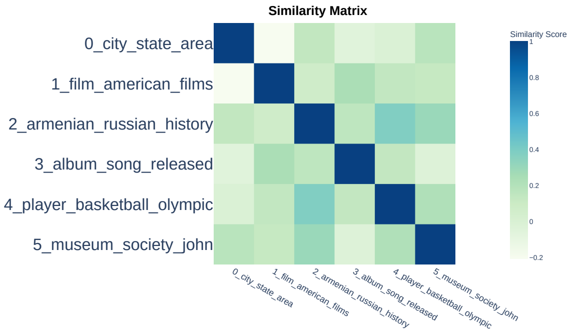


Figure 3: The similarity heatmap of the top 6 frequent topics present within the mined English paragraphs.

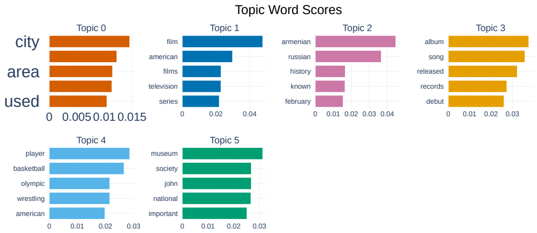


Figure 4: The usage of frequent words in the top 6 frequent topics present within the mined English paragraphs.

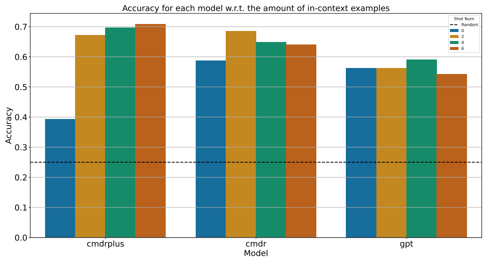


Figure 5: Accuracy of each model with a varying number of in-context examples given before generation.

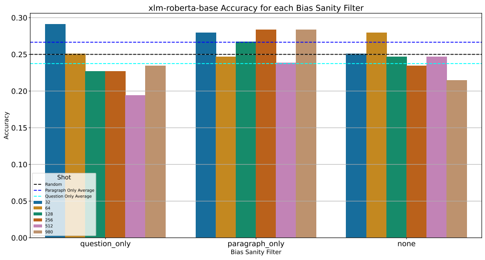


Figure 6: The results of fine-tuning XLM-Roberta on the Armenian QA dataset with a varying number of training samples while using only paragraphs, questions or random data.

A Appendix

Generated Question-Answer pairs We showcase examples of generated and validated question-answer pairs along with their designated English paragraph $\mathcal{P}_{\text{Eng}}$ in Table 6. These are representative samples of the generation process, further reinforced by the fact that human evaluation of the quality of the generation showed that 98% of the examples are answerable and maintain quality.

What are the questions about ? To understand the type of inquiries asked within the questions, we employ a pre-trained model for Named Entity Recognition (NER) from spaCy[1] and detect all the entity types mentioned within the question-answer pairs. The results can be seen in Table 5, showing that the object of the inquiries can vary massively from people (PERSON) and locations (LOC) to organization (ORG), numeric values (DATE, ORDINAL, TIME), etc. This further ensures that we are able to generate high-quality questions with diverse compositions and object of inquiry types.

Topic Distribution the parallel paragraphs To estimate the overlap within the topics found in the mined paragraphs, we use unsupervised topic modelling BERTopic (Grootendorst, 2022) to segment the 5 most frequently occurring segments. We measure the overlap between these by calculating the averaged cosine distance of the topic embeddings obtained from BERTopic. The results can be seen in Fig. 3 and Fig. 4, validating our hypothesis that we are able to cover diverse themes using our parallel paragraph mining method.

Benchmarking with Armenian QA dataset To show the usefulness of the created dataset, we benchmark several SOTA LLMs on it in supervised fine-tuning, *zero-shot* and *few-shot* settings. We further investigate if the dataset suffers from statistical biases or degenerate solutions by training an XLM-RoBERTa model on inputs that contain only the paragraph or the question, excluding everything else from the sample. The results in Fig. 6 show us that regardless of the amount of provided training samples, the question and paragraph-only evaluations behave similarly to random chance, highlighting that the dataset is unlikely to suffer from inconsistencies and degenerate solutions.

We benchmark several LLMs, shown in Fig. 5, using produced Armenian QA benchmark and show that while increasing the number of model pa-

[1] https://spacy.io/api/entityrecognizer

OTHER	NORP	GPE	PERCENT	PERSON	DATE	ORG	WORK OF ART	LANGUAGE	QUANTITY	EVENT	MONEY	LOC	ORDINAL	TIME	FAC	PRODUCT
3178	172	223	8	397	335	327	14	10	25	21	9	52	38	9	9	3

Table 5: Distribution of Entities within question-answer pairs in the generated English QA dataset. The Entity labelling scheme follows Honnibal et al.

Example 1: UEFA Champions League
Since the rebranding of the European Champion Clubs' Cup as the UEFA Champions League in 1992, 107 different players from 37 countries have scored three goals or more in a single match (a hat-trick) on 152 occasions, representing 53 clubs from 17 leagues. The first player to achieve the feat was Juul Ellerman, who scored three times for PSV Eindhoven in a 6–0 victory over Žalgiris on 16 September 1992. Lionel Messi and Cristiano Ronaldo have scored three or more goals in a match eight times each in the Champions League, more than any other player, followed by Robert Lewandowski with six, and Karim Benzema with four. **Question:** What was the original name of the UEFA Champions League? **Answers:** 1. European Champion Clubs' Cup, 2. European Premier League, 3. UEFA Football Cup, 4. European Soccer Championship **Correct Answer:** 1. European Champion Clubs' Cup
Example 2: Sign Languages
Sign languages (also known as signed languages) are languages that use the visual-manual modality to convey meaning, instead of spoken words. Sign languages are expressed through manual articulation in combination with non-manual markers. Sign languages are full-fledged natural languages with their own grammar and lexicon. Sign languages are not universal and are usually not mutually intelligible, although there are also similarities among different sign languages. **Question:** What is the primary modality used to convey meaning in sign languages? **Answers:** 1. Auditory-vocal, 2. Visual-manual, 3. Tactile-kinesthetic, 4. Olfactory-gustatory **Correct Answer:** 2. Visual-manual

Table 6: Examples of English paragraphs along with their generated question-answer pairs

rameters and in-context samples helps the overall model performance, still even very large models are unable to solve the dataset trivially, thus showing its value as a benchmarking resource.