Evaluation of Question Answer Generation for Portuguese: Insights and Datasets

Felipe S. F. Paula
Institute of Informatics
UFRGS
felipesfpaula@gmail.com

Cassiana R. L. Michelin
Institute of Geosciences
UFRGS
cassiana.michelin@ufrgs.br

Viviane P. Moreira
Institute of Informatics
UFRGS
viviane@inf.ufrgs.br

Abstract

Automatic question generation is an increasingly important task that can be applied in different settings, including educational purposes, data augmentation for question-answering (QA), and conversational systems. More specifically, we focus on question answer generation (QAG), which produces question-answer pairs given an input context. We adapt and apply QAG approaches to generate question-answer pairs for different domains and assess their capacity to generate accurate, diverse, and abundant question-answer pairs. Our analyses combine both qualitative and quantitative evaluations that allow insights into the quality and types of errors made by QAG methods. We also look into strategies for error filtering and their effects. Our work concentrates on Portuguese, a widely spoken language that is underrepresented in natural language processing research. To address the pressing need for resources, we generate and make available human-curated extractive QA datasets in three diverse domains.

1 Introduction

Over the years, question generation (QG) established itself as an important task in Natural Language Generation. QG has been used in asking clarification questions in dialogue (Majumder et al., 2021) and for providing follow-up questions after an answer (Meng et al., 2023). Also, question generation probabilities can be used to improve information retrieval (Sachan et al., 2022). It is especially important in educational settings, where it can be used to generate quizzes automatically (Laban et al., 2022). In addition, it can be used for data augmentation aiming at domain adaptation (Shakeri et al., 2020) or enhancing QG robustness (Bartolo et al., 2021).

Questions can be generated in an answer-aware or answer-agnostic fashion. Question-answer generation (QAG) (Ushio et al., 2023) is the task that aims at producing a set of questions along with their respective *extractive* answers, taking a context passage as input. QAG is more complex than general QG since it needs to produce not only the questions but also the answers.

Portuguese is the sixth largest language in terms of the number of native speakers and yet it is underrepresented in terms of linguistic resources. To the best of our knowledge, there is only one fully extractive QA dataset, namely FaQuAD (Sayama et al., 2019), which contains 900 question-answer pairs. Thus, it is pressing that more resources be created for this language.

Although much has been done in the field of QG, some gaps remain. Current studies paid little attention to the decoding method, which is an important factor regarding the resulting accuracy, diversity, and adequacy of the output. Also, some works focus on the downstream performance of generated questions in QA tasks, not assessing whether the generated questions are correct (Bartolo et al., 2021; Ushio et al., 2023; Shakeri et al., 2020).

In this work, we adapt and apply two QAG approaches to generate question-answer pairs for three domains in Portuguese using different decoding methods. Our contributions include:

1. Quantitative and qualitative analyses of QAG approaches applied to an underrepresented language. We assess their capacity to generate accurate, diverse, and abundant question-answer pairs under different settings.
2. An evaluation of question filtering methods gauging what kind of errors they prevent, their effects on question distribution, and their efficiency in selecting accurate questions.
3. QAG datasets in three diverse domains containing 993 question-answer pairs that were validated by humans. The datasets are available at this link[1].

[1] https://github.com/felipesfpaula/qgen_submission

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 5315–5327

The main findings from our analyses can be summarized as follows:

- Beam search decoding generates more accurate questions despite being less diverse.
- The end-to-end method (Ushio et al., 2023), despite being useful to QA data augmentation, generates more questions disconnected from their answer spans.
- The types of question-answer pairs produced depend upon the kind of available information in the input text and the generation bias of the models.
- In some cases, filtering using the answer score probabilities may lead to discarding questions with more complex answers.
- Common question filtering methods, like Roundtrip consistency, while preventing answerability and coherence errors, do not noticeably mitigate fluency errors.
- Filtering methods using an ensemble of other filters combined in a relaxed voting scheme can be selective and efficient.

2 Related Work

Question Generation. Early systems were based on rules or templates that transformed declarative structures using the available syntactic or semantic information (Zhang et al., 2021). With the advent of sequence-to-sequence deep learning architectures (Sutskever et al., 2014) and the availability of large-scale question datasets, such as SQuAD (Rajpurkar et al., 2016), the paradigm shifted towards data-driven approaches (Du et al., 2017; Du and Cardie, 2018; Zhou et al., 2018). More recently, with the success of pre-trained language models (PLMs) in many tasks (Radford et al., 2018; Devlin et al., 2019; Raffel et al., 2020), many solutions approached question generation by fine-tuning these models. As the latest development, large language models (LLMs), such as ChatGPT, present In-Context Learning capabilities that allow them to adapt themselves to a new task with few examples and without explicit training. For instance, LLMs have been utilized to generate questions that serve in evaluating Retrieval-Augmented Generation (RAG) systems (Chen et al., 2024).

Question Generation Evaluation. When ground truth answers are available, the generated questions can be evaluated with respect to them. This is done with metrics that quantify lexical matching, such as BLEU(Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), and ROUGE (Lin, 2004), or metrics that assess semantic similarity, like BERTScore (Zhang et al., 2020). Although widely used in the literature, these metrics have important limitations. BERTScore and BLEU may yield high scores to questions with swapped entities (Mohammadshahi et al., 2023), for example. Moreover, reference-based metrics correlate poorly with human raters, although this can be ameliorated with more reference questions (Oh et al., 2023). To address this problem, reference-free metrics that use PLMs to assess question adequacy were proposed like QRelScore (Wang et al., 2022) and RQUGE (Mohammadshahi et al., 2023).

Question Generation in Portuguese. Most of the work in QG for Portuguese has been rule-based. In one of the earliest studies, Diéguez et al. (2011) combined case-based reasoning techniques with rule-based approaches to transform declarative sentences into questions. Pirovani et al. (2017) utilized named-entity recognition to identify possible answers and employed this information to generate close-type and discursive questions using rules. Some works have compared the performance of rule-based methods that leverage syntactic, semantic, and dependency information (Leite et al., 2020; Ferreira et al., 2020; Leite and Cardoso, 2023). Until recently, the performance of data-driven QG methods had not been thoroughly assessed in Portuguese. Leite and Lopes Cardoso (2022) conducted a pilot study on the performance of PLM-based generation and found that it was comparable to results in English.

The work by Ushio et al. (2023) is related to ours as we compare similar QAG strategies across different domains. However, unlike these authors, we also assess the qualitative aspects of QAG using human evaluation. Bartolo et al. (2021) also assessed the performance of different question filtering approaches. In this work, we carried out a deeper analysis by looking into decoding strategies and additional filtering methods.

3 Materials and Methods

Most current solutions for QAG either employ PLMs with fine-tuning (Ushio et al., 2023) or LLM with prompting. Both have been applied to different languages and domains. Our goal is to compare their performance according to quantitative

and qualitative criteria across different domains in Portuguese.

To serve as training data, we translated SQuAD 1.1 (which is licensed under CC BY-SA 4.0) from English (Rajpurkar et al., 2016) to Portuguese using the GoogleTranslate API[2]. Despite the availability of other Portuguese translations of SQuAD v1.1, we opted to create our own version to ensure precise alignment of the answer span offsets between the answers and the contexts. Maintaining this alignment is crucial because misaligned offsets in existing translations would introduce mismatches between answers and contexts, leading to the loss of training instances during model training. We will refer to our version as PT-SQuAD and make it available at our repository[3]. More details on this translation process are in Appendix A. A visual inspection of a sample of the translated contexts and questions showed that the translation quality was fair.

3.1 QAG Methods

We follow Ushio et al. (2023) and tested two relevant QAG methods based on PLMs: *end-to-end* and *pipeline*.

End-to-end (E2E). Given an input paragraph, this approach generates a sequence of questions and answers in the format "question:{q_1}, answer:{a_1} | question:{q_2}, answer:{a_2} | ... ". For example:

```
Input: Dengue fever is a mosquito-
borne disease caused by dengue
virus, prevalent in tropical and
subtropical areas.
Output: question: {What is the
mosquito-borne disease caused by
the dengue virus?}, answer:{Dengue
fever} | question: {Where is
dengue fever prevalent?}, an-
swer:{in tropical and subtropical
areas}
```

Pipeline (Pipe). This method consists of two sequential steps: answer extraction and answer-aware question generation. We employed Self-Attention Labeling (SAL) (Bartolo et al., 2021) as the answer extraction method. SAL uses the last hidden layer of an encoder-based PLM to model the start and end vectors of a possible answer span. These vector representations, **S** (start) and **E** (end), are the *keys* and *queries* for an operation similar to the scaled dot-product attention. The probability of span (i, j) being an answer in context c is given by:

$$p(a_{ij}|c) = \sigma\left(\frac{(W_1\mathbf{S})(W_2\mathbf{E})^\top}{\sqrt{d}}\right) \quad (1)$$

where $\sigma(\cdot)$ is the sigmoid function and d the hidden layer dimension. Unlike Bartolo et al. (2021), we add two feature extraction matrices W_1 and W_2. The intuition was to enhance representation learning and bring the formula closer to the multi-headed attention model. BERTimbau-Large (Souza et al., 2020), a Portuguese version of BERT, was used as the encoder. Once the candidate answer is extracted by SAL, we need to "highlight" it in the context c to generate each question. This is done by adding a special token <hl>.

```
Input: <hl> Dengue fever <hl> is
a mosquito-borne disease caused by
dengue virus, prevalent in tropical
and subtropical areas.
Output: What is the mosquito-borne
disease caused by the dengue virus?
```

Question Generation via Prompting We use the in-context-learning capabilities of LLMs to generate questions in a few-shot manner. PT-SQuAD was used as a source of examples. More specifically, given an example context and its respective questions from PT-SQuAD, we asked the LLM to generate five questions along with their answers in the form of verbatim spans of the input paragraph. Preliminary tests revealed that if the prompt did not contain the number of questions to be generated, then the LLMs would generate just a couple. On the other hand, if we asked for too many, then quality degraded noticeably. As a result, five was seen as an adequate choice. We tested the models GPT-3.5-turbo[4] from OpenAI and Sabiá-medium (Pires et al., 2023) from MaritacaAI[5], a model specifically built for Portuguese. The complete prompt is in Appendix B.

3.2 Decoding Methods

Beam search This decoding method searches for the highest probable sequence output by keeping a number of subsequence hypotheses, called beams. However, despite the outputs being assigned a high

[2] https://translate.google.com/

[3] https://github.com/felipesfpaula/qgen_submission

[4] The choice for GPT-3.5-turbo was motivated by budget restrictions

[5] https://www.maritaca.ai/

probability, the generated text may sound artificial, awkward, and prone to repetitions (Fan et al., 2018; Holtzman et al., 2020). Nevertheless, it can yield good results when the length of the desired generation is somewhat predictable. We choose to decode our outputs with five beams.

Top-p nucleus sampling (Holtzman et al., 2020) This decoding strategy samples from the highest probability tokens whose cumulative probability mass exceeds the threshold p. By sampling these high-probable tokens (the nucleus), top-p avoids selecting less probable tokens that can degenerate the output. In our experiments, we follow Bartolo et al. (2021) and set $p = 0.75$.

3.3 Domain Corpora

To evaluate QAG approaches, we selected input texts from different domains shown in Table 1. The pipeline and end-to-end methods were fine-tuned on PT-SQuAD, which is based on Wikipedia-EN. Then, to have a corpus on a related domain, we selected some passages from Wikipedia-PT. To avoid overlaps with the training data, we manually picked passages from pages that were not in SQuAD. The intuition is that QAG methods would perform better in corpora that are closer to its training data. For an alternative non-technical domain, we collected articles from a few Brazilian newspapers published in March 2024 to ensure the contents were not in the training data of the LLMs.

Additionally, to gauge QAG performance within a technical domain, which would likely be more challenging, we resorted to the Geoscientific domain. This domain is relevant for Portuguese-speaking countries, given the economic importance of the Oil & Gas industry, which represents a significant portion of these countries' GDP (57% in Angola and 13% in Brazil). Yet, there are not many linguistic resources available for specific domains in under-resourced languages. One of the few resources is the REGIS collection(Lima de Oliveira et al., 2021) (available under the MIT license) that was used here. Documents in this collection typically focus on Geoscientific research and exploration activities in the oil and gas industry. They often include scientific studies, research findings, technical reports, and data analyses related to geological, geophysical, and geochemical aspects of oil and gas exploration and production.

From each of the three corpora, we selected 15 passages containing 384 BERTimbau tokens each. These passages were used as contexts for generating questions in the experiments reported in the next sections.

Corpus	Description	#Questions
Wiki	Passages sourced from pt-Wikipedia	692
News	Passages collected from the news on March 2024	599
geoREGIS	Passages selected from random paragraphs of geoscientific documents within REGIS.	528

Table 1: Overview of corpora used as input to the QG and the total output of questions

3.4 Models and Training

Our choice of model was influenced by the good results achieved by T5 (Raffel et al., 2020) in generating questions that improve QA quality in English (Ushio et al., 2023). For Portuguese, Leite and Lopes Cardoso (2022) showed that the PTT5 (Carmo et al., 2020) was better at generating questions than multilingual T5 (mT5) (Xue et al., 2021). For this reason, we chose PTT5 as the backbone of our QAG methods. Finally, we fine-tuned the PTT5 models on the 87K instances of PT-SQuAD. Further implementation details are in Appendix D.

4 Evaluating Question Quality

To assess the quality of the questions produced by the different QAG methods and decoding strategies, we used the error taxonomy proposed by Laban et al. (2022). Table 2 presents the error classes. Errors are classified into two granularity levels. At the macro level, the question can be rejected because it is *disfluent* (problems in grammar or phrasing), *off-target* (problems in answerability), and *wrong context* (problems in consistency and specificity). At the micro level, there are ten classes that further specialize in the types of problems the questions could incur.

Annotation. Human annotators classified the automatically generated questions according to the error taxonomy. We recruited five native Portuguese speakers from our research group (two females and three males aged between mid-20s and early 40s). They were compensated at an hourly rate that was well above the regional minimum wage. The annotators were provided with guidelines and examples and were able to discuss them with the

authors. The complete guidelines are in our repository[6] and a screenshot of the annotation interface is in Appendix C. For the geoscientific domain, which demands specific knowledge, one annotator had a PhD in geology, and the other one was a third-year BSc student. The inter-annotator agreement calculated using Krippendorff's α was 0.49, which we consider a fair agreement.

Quality Metrics. For each generation model and decoding strategy, we computed three evaluation metrics that capture the desired properties of QAG. *Accuracy* (Acc) is the percentage of correct questions generated by the setup. *Productivity* (Prod) is the mean number of question-answer pairs generated by each context. *Diversity* (Div) is $1-b$, where b is the self-BLEU measure (Zhu et al., 2018) of the generated questions by the setup, and it is higher when there is little overlap among the generated questions.

Six different setups were evaluated for QAG in each of the three domains. The results for the evaluation metrics are shown in Table 3, and the distribution of the types of errors for each setup are in Figure 1.

E2E vs. Pipe. The pipeline method generates more correct questions than end-to-end. In Wikipedia and the Geoscientific domain, the pipeline is more productive of question-answer pairs. The pipeline method presents the same productivity for beam search and top-p because they use the same answers as input and thus generate the same number of questions. It is not clear which method produces a more diverse set of questions. End-to-end models present more *off-Target* errors compared to pipeline. We noticed that E2E has more trouble selecting answers. The pipeline method has a more accurate answer selection method, and as a result, it commits fewer *off-Target* errors. Since this error class corresponds to mismatches between the answer span and the question, we can see that compared to the pipeline, E2E has more trouble forming connections between the answer and the question, highlighting the characteristic of not being conditioned on the answer.

Beam Search vs. Top-p Regarding the decoding strategy, beam search was consistently more accurate than top-p nucleus sampling for both generation strategies and across all domains. On the other hand, top-p sampling shows more diversity. In the end-to-end setting, the beam search decoding produces more questions.

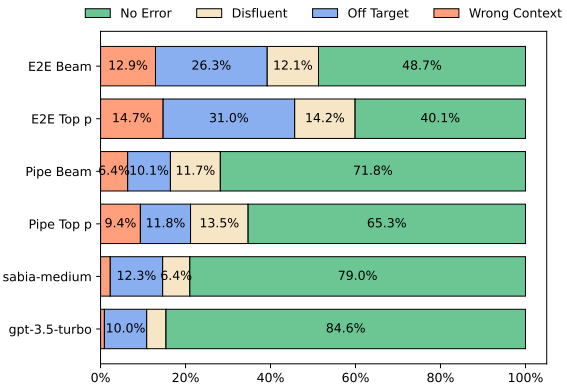

Figure 1: Distribution of errors for the different setups. The scores are for all domains combined.

PLM vs LLM. The high scores achieved by Sabiá-medium and GPT-3.5-turbo convey the good performance of bigger, more capable models, which can be achieved even in a few-shot manner. They are more accurate and more diverse than the PLM-based models. Their main errors in LLM-based QAG come from the *off-Target* class. A possible reason for that is that we ask them to generate answer spans, and in their non-deterministic decoding process, they could change some tokens, causing a mismatch between the span in the text and their answer.

Question error classes. The results of the human annotation of the error classes are shown in Figure 1. A general tendency across all setups is that the most frequent error class was off-target. Comparing the PLM-based models, E2E is worse than pipeline in the three error classes, although they are comparable in fluency. Many of the errors of E2E come from the *off-Target* class. Although the distribution of the coarse-grained error classes looks similar for E2E beam search and top-p nucleus sampling, they are different for the fine-grained ones.

5 Evaluating Answers

Looking into the answers provides strong cues about the questions. To identify what questions are being asked by the QAG strategies, we devised an answer classification scheme. This taxonomy aims to identify biases in the type of answers and their interdependence with the domain.

[6] https://github.com/felipesfpaula/qgen_submission

Error Category	Fine-grained	Examples
Disfluent	Wrong Tense	Q: Qual é a palavra que Freud usou para descrever sua doença?
	Awkward Phrasing	A: neurose
	Not a Question	*Q: What is the word Freud used to describe his disease?*
	Repetitions	*A: neurosis*
Off-Target	Unanswerable	Q: Qual é a massa do próton?
		A: 1 / 1836 da massa do próton
	Other answer span	Q: *What is the mass of proton?*
		A:*1 / 1836 off the mass of proton*
Wrong Context	Too specific	Q: Qual é o nome do campo de Namorado?
	Reveals answer	A: Namorado
	Inconsistent	Q: *What is the name of the field Namorado?*
	Not specific enough	A: *Namorado*

Table 2: Error taxonomy with examples of questions and answers. Translations into English are provided for clarity.

		Wiki			**News**			**GeoREGIS**		
		Acc	*Prod*	*Div*	*Acc*	*Prod*	*Div*	*Acc*	*Prod*	*Div*
E2E	beam	0.43	5.80	0.46	0.44	6.46	0.47	0.68	3.38	0.77
	top-p	0.37	5.40	0.71	0.38	5.73	0.67	0.48	2.45	0.91
Pipe	beam	0.73	7.93	0.55	0.65	5.86	0.65	0.74	6.07	0.68
	top-p	0.64	7.93	0.68	0.56	5.86	0.76	0.72	5.61	0.74
	Sabiá	0.68	5.00	0.73	0.88	5.00	0.76	0.78	4.58	0.77
	GPT-3.5-turbo	0.80	5.00	0.77	0.93	5.00	0.78	0.79	4.83	0.73

Table 3: Assessment of the QAG methods in terms of accuracy, productivity, and diversity.

While there are other question classification taxonomies like Bloom's Cognitive Levels (Bloom et al., 1956), TREC-10 (Li and Roth, 2002), and Graesser's (Graesser et al., 2008), we did not find them suitable to our goals. We have verified a low adherence to Bloom's and Graesser's taxonomy classes toward PT-SQuAD data. Consequently, we adopted a classification scheme that simplifies the TREC-10 taxonomy.

We manually classified answer spans from correct question-answer pairs selected in the previous experiment. Our classification scheme for types of answers includes the following four classes.

- **Name**. Names or lists of names of people, locations, organizations, creative works, *etc*.
- **Numeric**. Dates, currency, measurements, and quantitative information in general. The format can be textual or numeric.
- **Description**. Longer textual forms that could function as a sentence.
- **Term**. Single words, multi-word expressions, or lists of words that are not proper nouns or form a sentence.

Figure 2 shows the results of these analyses. To use as a reference, we also manually annotated 414 ground truth answers from PT-SQuAD. With these questions, we can have a sample of humans' extractive answers to questions in the Wikipedia domain.

Despite the increased focus on the *Term* class, we found that the E2E and Pipe answer distributions are compatible with PT-SQuAD. A possible reason for the difference is that our Wikipedia-PT subsample contains more salient terms. In the news domain, all generation strategies focused more on names and less on terms. This agrees with the intuition that news articles feature more discussion about named entities than debates about concepts. Text from the geoscientific domain presents a fair amount of names of geological formations, basins, and locations. Across the three domains, the LLM approach, GPT-3.5-turbo and Sabiá-medium, is biased towards longer and more descriptive answers. Overall, the data suggests that the question-answer distribution is defined by the interaction between the types of question-answer pairs the methods tend to produce and the available salient information in the source passages.

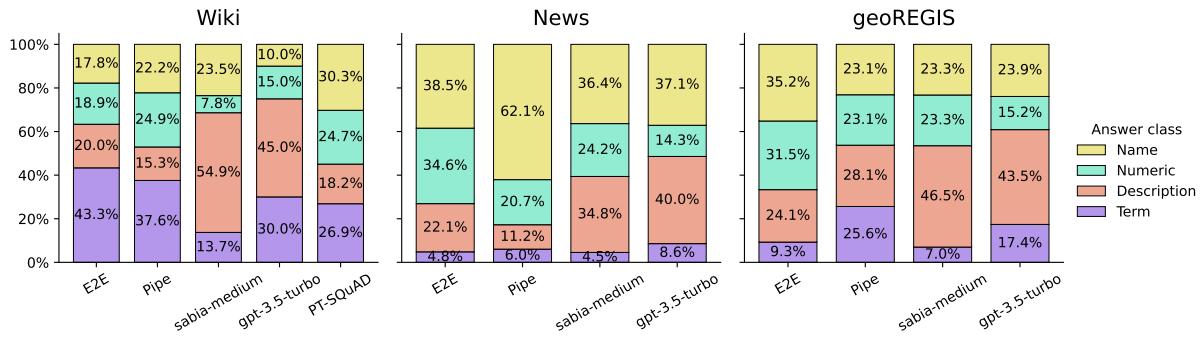

Figure 2: Answer class distribution per domain.

6 Question Filtering

Filtering methods can be used to reduce noisy and erroneous questions. We focus on assessing the effects of the filtering methods on question distribution and what errors the filtering prevents. In the following analysis we consider only the Pipeline method. We implemented filtering methods based on the following criteria.

Answer Score (Bartolo et al., 2021). This filter selects questions in which the probability output of the answer candidate defined in Equation 1 surpasses a given threshold k.

Generation Score (GS) (Bartolo et al., 2021). This filtering strategy selects questions with generation probability above a threshold k. In other words, the probability of the sequence of tokens that constitutes the question must surpass a threshold.

Roundtrip (Alberti et al., 2019; Bartolo et al., 2021). For each question candidate and respective context pair, we check if the candidate answer is the same as the answer given by the 6-way QA ensemble model with the same question and context. We fine-tuned six different BERTimbau QA models using different random seeds on PT-SQuAD data. We select candidates for which the answer agrees with at least one QA answer, case "1/6", and when it agrees with all six, case "6/6".

QRelScore LRM (Wang et al., 2022). The reference-free metric QRelScore can also be used to filter questions. Originally, this metric used RoBERTa and GPT-2 as components to calculate two sub-metrics: local relevance matching (LRM) and global relevance generation (GRG), respectively. However, we switched RoBERTa (Liu et al., 2019) to BERTimbau since our data is in Portuguese. We also tried switching GPT-2 (Radford et al., 2019) to XGLM (Conneau and Lample, 2019). However, it yielded poor GRG results and hurt the QRelScore performance. Thus, we report only the LRM values. We filter candidates by selecting instances with the QRelScore LRM above the sample mean.

ϕ voting. To combine the strengths of previous filtering methods, we establish two voting schemes using the Generation Score > 0.9, Roundtrip 1/6, and QRelScore LRM. The first one, *relaxed*, accepts the candidate if one of the conditions is met. The second one, *strict*, only accepts a candidate if the three conditions are met.

The answer score does not select correct questions across different domains. In fact, it only helps to select the right questions in the Wikipedia domain, which is the same domain as the training data (see Appendix E). The proportion of *description* type of questions gets lower by using higher thresholds of answer score. This means there are fewer questions that target answers that require more elaborate descriptions. In Figure 3, we plot the description class proportion by the generation score metric and the answer score. The answer span thresholds select simpler answers with less structured text. This does not necessarily mean shorter answers. As a result, we omit it from the next analyses.

The errors filtering helps to prevent can be seen in Figure 4. The generation score thresholds reduce the *wrong context* error class and slightly improve the *off-target* class. Roundtrip in both versions greatly reduces the *off-target* and slightly decreases the *wrong context* class. The voting strategy ϕ_{relaxed} shows performance comparable to Roundtrip 1/6 and ϕ_{strict} only presents *disfluent* errors. QRelScore LRM is a marginal improvement over unfiltered question candidates. The results

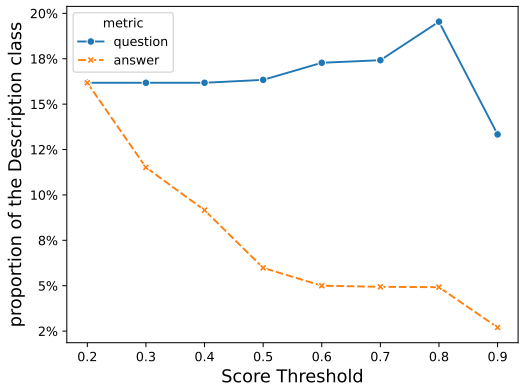

Figure 3: Proportion of the description class according to the answer and generation scores.

showed that fluency errors resist current filtering strategies. Roundtrip consistency and QRelScore LRM consider the passage-question pair coherence through the answer and semantic similarity, respectively. The generation score thresholds could improve fluency. However, it actually improves the consistency and adequacy of questions.

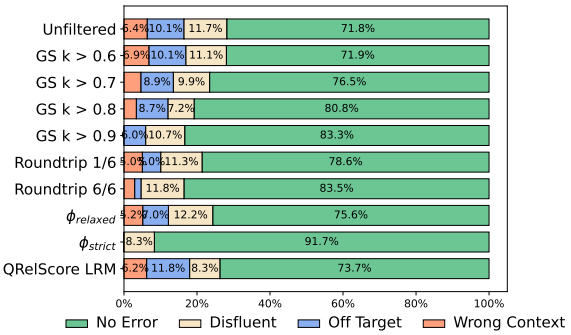

Figure 4: Error distribution of filtering strategies. The bars tagged with k represent the generation score at different thresholds.

To evaluate how well the filtering techniques handle the relevant (correct) questions, we measure the precision, recall, and f1. A high precision score means that the filtering technique effectively excludes incorrect questions. On the other hand, a high recall means that the filter does not discard correct questions. We postulate that the filtering process should strive to reach a balance between precision and recall. In Table 4, we can see the performance of the filtering methods classifying correct questions per domain. The unfiltered questions present a baseline across the domains. For Wikipedia and Geoscientific domains, QRelScore LRM presents the best precision. However, it is moderately low in the News domain. Roundtrip 1/6 shows a good balance between precision and recall, achieving the best F1 score in the News domain. On the other hand, Roundtrip 6/6 is slightly more precision-oriented. The strict voting strategy ϕ_{strict} achieves perfect precision on Wikipedia and News domains, although it shows a negligible recall across the three domains. Finally, the relaxed voting strategy ϕ_{relaxed} also shows a good trade-off of precision and recall. ϕ_{relaxed} shows the best recall of all filtering methods studied here.

Although filtering should focus on precision, recall is also important. Applications of QAG may also need a volume of questions, and a highly selective filtering method may not work. For this purpose, an ensemble of filtering methods in a *relaxed* voting scheme may be more appropriate.

7 Conclusion

In this study, we provided an in-depth examination of the quality of the generated question-answer pairs. Our goal was to identify the strengths and weaknesses of current QAG methods. We investigated the impact of different decoding methods, assessed the accuracy and diversity of generated question-answer pairs, and analyzed the types of errors made by various QAG models. Additionally, we explored the classification of questions based on the model output and evaluated the effectiveness of different question filtering methods.

The preference for beam search decoding in QAG systems may need to be reconsidered based on the specific application. While it ensures higher accuracy, the reduction in diversity could limit the range of questions generated, which might be crucial in educational settings or for generating varied training data. In educational technology, the findings suggest that QAG tools must be carefully designed to balance accuracy and diversity to create comprehensive assessments and learning materials. For training QA systems, ensuring that the generated questions are both accurate and varied can enhance robustness. The insights from this study could guide the development of more effective training datasets. Regarding the influence of the source context in QAG, the dependency on the input context suggests that QAG systems need to be tailored to handle various types of source material.

The study of accuracy in generation can be expanded to include non-extractive questions. Despite the importance, QAG with generative answers

	Wikipedia			News			Geoscientific		
	Prec	*Rec*	*F1*	*Prec*	*Rec*	*F1*	*Prec*	*Rec*	*F1*
Unfiltered	0.73	1.00	0.84	0.65	1.00	0.79	0.75	1.00	**0.86**
QRelScore LRM	0.85	0.58	0.69	0.58	0.43	0.49	0.85	0.59	0.70
Roundtrip 1/6	0.80	0.85	0.82	0.75	0.96	**0.84**	0.79	0.82	0.80
Roundtrip 6/6	0.84	0.60	0.70	0.83	0.77	0.80	0.83	0.63	0.72
ϕ_{strict}	1.00	0.08	0.14	1.00	0.01	0.03	0.75	0.04	0.08
ϕ_{relaxed}	0.80	0.97	**0.88**	0.69	0.98	0.81	0.78	0.89	0.83

Table 4: Classification performance metrics for correct questions.

is still a poorly explored subject. Furthermore, current question filtering methods are too precision-oriented. New filtering strategies that also present a high recall need to be developed so QAG applications can benefit from more questions. Finally, current filtering methods do not mitigate fluency errors. A possible future inquiry may be whether filtering based on linguistic acceptability (Warstadt et al., 2019) improves performance.

8 Limitations

Our findings are based on the analysis of QAG in a single language. To verify how well our results generalize, this study may need to be replicated in different languages.

QAG methods are able to generate a very large number of question-answer pairs and thus produce very large datasets. However, in this work, human curation and manual error classification meant that the datasets we made available are small. Nevertheless, they are larger than the only previously existing dataset.

Additionally, the input data of the PLM-based QAG methods examined in this paper resulted from the automatic translation of data originally in English. While language can be translated with good accuracy, the data may not represent the cultural aspects of the target language.

Finally, although we experimented with different decoding methods, we did not vary their parameters systematically to gauge their impact on the output question-answer pairs.

9 Ethical Considerations

This work relied on human annotators who were compensated fairly, receiving payments that were well above the minimum wage for our region. We worked with data from Wikipedia, news articles, and geoscientific reports. Questions and answers did not include contents that could potentially be considered offensive or upsetting. We used LLMs for question generation to serve as a basis for comparison with PLM approaches. LLM-generated data may contain pernicious biases. Nevertheless, the human annotators did not identify anything in particular that was worthy of flagging in this respect.

Acknowledgments

This work has been partially funded by CENPES Petrobras, CNPq-Brazil, and Capes Finance Code 001.

References

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173, Florence, Italy. Association for Computational Linguistics.

Max Bartolo, Tristan Thrush, Robin Jia, Sebastian Riedel, Pontus Stenetorp, and Douwe Kiela. 2021. Improving question answering model robustness with synthetic adversarial data generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8830–8848, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Benjamin S Bloom, Max D Engelhart, EJ Furst, Walker H Hill, and David R Krathwohl. 1956. Handbook i: cognitive domain. *New York: David McKay*.

Diedre Carmo, Marcos Piau, Israel Campiotti, Rodrigo Frassetto Nogueira, and Roberto de Alencar Lotufo. 2020. PTT5: pretraining and validating the T5 model on brazilian portuguese data. *CoRR*, abs/2008.09144.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in

retrieval-augmented generation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17754–17762. AAAI Press.

Alexis Conneau and Guillaume Lample. 2019. *Cross-lingual language model pretraining*. Curran Associates Inc., Red Hook, NY, USA.

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.

Daniel Diéguez, Ricardo Melo, and Paulo Gomes. 2011. Using cbr for portuguese question generation. In *Progress on artificial intelligence: 15th Portuguese Conference on Artificial Intelligence*, 15, pages 328–341.

Xinya Du and Claire Cardie. 2018. Harvesting paragraph-level question-answer pairs from Wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1907–1917, Melbourne, Australia. Association for Computational Linguistics.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

João Ferreira, Ricardo Rodrigues, and Hugo Gonçalo Oliveira. 2020. Assessing Factoid Question-Answer Generation for Portuguese. In *9th Symposium on Languages, Applications and Technologies (SLATE 2020)*, volume 83 of *Open Access Series in Informatics (OASIcs)*, pages 16:1–16:9, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Art Graesser, Vasile Rus, and Zhiqiang Cai. 2008. Question classification schemes. In *Proc. of the Workshop on Question Generation*, pages 10–17.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Philippe Laban, Chien-Sheng Wu, Lidiya Murakhovs'ka, Wenhao Liu, and Caiming Xiong. 2022. Quiz design task: Helping teachers create quizzes with automated question generation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 102–111, Seattle, United States. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Bernardo Leite and Henrique Cardoso. 2023. Do rules still rule? comprehensive evaluation of a rule-based question generation system. In *Proceedings of the 15th International Conference on Computer Supported Education*, page 27–38. SCITEPRESS - Science and Technology Publications.

Bernardo Leite, Henrique Lopes Cardoso, Luís Paulo Reis, and Carlos Soares. 2020. Factual question generation for the portuguese language. In *2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, pages 1–7.

Bernardo Leite and Henrique Lopes Cardoso. 2022. Neural question generation for the portuguese language: A preliminary study. In *EPIA Conference on Artificial Intelligence*, pages 780–793. Springer.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Lucas Lima de Oliveira, Regis Kruel Romeu, and Viviane Pereira Moreira. 2021. Regis: A test collection for geoscientific documents in portuguese. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2363–2368, New York, NY, USA. Association for Computing Machinery.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Bodhisattwa Prasad Majumder, Sudha Rao, Michel Galley, and Julian McAuley. 2021. Ask what's missing and what's useful: Improving clarification question generation using global knowledge. In *Proceedings*

of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4300–4312, Online. Association for Computational Linguistics.

Yan Meng, Liangming Pan, Yixin Cao, and Min-Yen Kan. 2023. FollowupQG: Towards information-seeking follow-up question generation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 252–271, Nusa Dua, Bali. Association for Computational Linguistics.

Alireza Mohammadshahi, Thomas Scialom, Majid Yazdani, Pouya Yanki, Angela Fan, James Henderson, and Marzieh Saeidi. 2023. RQUGE: Reference-free metric for evaluating question generation by answering the question. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6845–6867, Toronto, Canada. Association for Computational Linguistics.

Shinhyeok Oh, Hyojun Go, Hyeongdon Moon, Yunsung Lee, Myeongho Jeong, Hyun Seung Lee, and Seungtaek Choi. 2023. Evaluation of question generation needs more references. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6358–6367, Toronto, Canada. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. *Sabiá: Portuguese Large Language Models*, page 226–240. Springer Nature Switzerland.

Juliana Pirovani, Marcos Spalenza, and Elias Oliveira. 2017. Geração automática de questões a partir do reconhecimento de entidades nomeadas em textos didáticos. In *Anais do XXVIII Simpósio Brasileiro de Informática na Educação (SBIE 2017)*, volume 1 of *SBIE*, page 1147. Brazilian Computer Society (Sociedade Brasileira de Computação - SBC).

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. Preprint.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Devendra Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3781–3797, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hélio Fonseca Sayama, Anderson Viçoso Araujo, and Eraldo Rezende Fernandes. 2019. FaQuAD: Reading comprehension dataset in the domain of brazilian higher education. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 443–448.

Siamak Shakeri, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Feng Nan, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. End-to-end synthetic data generation for domain adaptation of question answering systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5445–5460, Online. Association for Computational Linguistics.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. 2023. An empirical comparison of LM-based question and answer generation methods. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14262–14272, Toronto, Canada. Association for Computational Linguistics.

Xiaoqiang Wang, Bang Liu, Siliang Tang, and Lingfei Wu. 2022. QRelScore: Better evaluating generated questions with deeper understanding of context-aware relevance. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 562–581, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual

pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Ruqing Zhang, Jiafeng Guo, Lu Chen, Yixing Fan, and Xueqi Cheng. 2021. A review on question generation from natural language text. *ACM Trans. Inf. Syst.*, 40(1).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2018. Neural question generation from text: A preliminary study. In *Natural Language Processing and Chinese Computing*, pages 662–671, Cham. Springer International Publishing.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 1097–1100, New York, NY, USA. Association for Computing Machinery.

A SQuAD 1.1 Translation

Translating SQuAD poses challenges due to its structure, particularly in maintaining alignment between answer span boundaries and the context. Previous translations available online have not adequately addressed span alignment, resulting in the loss of some answers. In our approach, we annotate the answer spans in the contexts with pseudo-HTML tags ⟨/id/⟩ and ⟨//id/⟩, where "id" represents the identifier of the answer. The Google Translate API generally ignores HTML-like tokens, allowing us to reconstruct the spans corresponding to their respective answers accurately.

B Prompt

In our study, we tasked the LLMs with generating five questions and their verbatim span answers from a given PT-SQuAD context. By testing with different prompts, we found that without specifying the number of questions, the LLM generated only a few sample question-answer pairs, whereas requesting too many resulted in decreased quality. Consequently, five questions were deemed an optimal number.

```
You are a question generation bot.
The questions generated here will
be used as training data for another
bot.  Generate 5 pairs of extractive
questions and answers that must be
answered with only, and solely only,
with verbatim spans of the input text.
Examples:  {example} Input text:  {in-
put text}
```

Prompt 1: English translation of the input prompt submitted to *GPT-3.5-turbo* and *Sabiá-medium*

C Annotation System

We devised an interface to enable the annotation of the questions. A screenshot is shown in Figure 5.

D Implementation Details

The models were implemented using the Transformers library and run in Google Colab infrastructure. We can see the main hyperparameters in Table 5.

	E2E	Pipe	SAL
PLM	PTT5-Large	PTT5-Large	BERTimbau-large
Learning rate	1e-4	1e-4	2e-5
Batch size	4	4	16
Gradient Accumulation	8	8	0
Training Epochs	6	6	40

Table 5: Hyperparameters of PLMs used in this paper

The financial cost of the experiments was USD 20 for GPT experiments (OpenAI) and USD 4 for MaritacaAI.

It took around 8 hours to fine-tune the PLMs for QAG (Pipe and E2E combined). To fine-tune the BERTimbau-based QA ensemble also took 8 hours.

E Answer Score

In Figure 6, we can see that the answer score thresholds do not improve question accuracy across all domains. In the geoscientific domain, the answer score thresholds degrade the quality of questions.

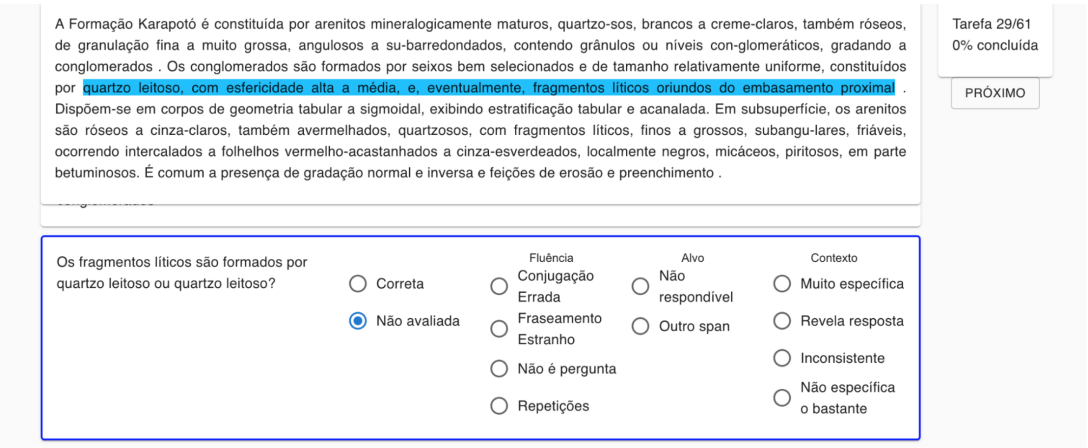

Figure 5: Screen capture of the annotation system

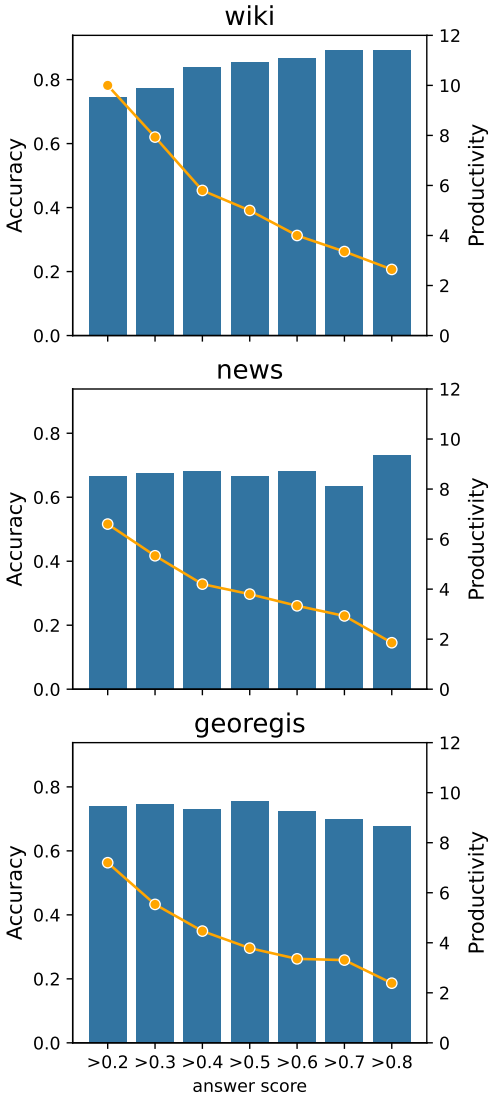

Figure 6: Percentage of correct questions at various thresholds for answer score.